MACHINE TRANSLATION

A Report for the BT8235_FinalProjectReport of Capstone Design -I

Yogesh Rathore
B-Tech / CSE
Galgotias University
Noida, India
Yrathore156@gmail.com

Md Farman
B-Tech / CSE
Galgotias University
Noida, India
Farman0150@outlook.com

Dhiraj Kumar Gupta
B-Tech / CSE
Galgotias University
Noida, India
Kumardhiraj424@gmail.com

Abstract — Language is an important means of communication and translation is a basic tool for understanding information in an unfamiliar language. Machine translation helps people to understand the details of an unknown language without the help of a human translator. This study is about to translate one language into another.

I. LITERATURE REVIEWS

MT systems can be categorized according to their translation methods. The classification of MT methods is given in Figure 1. MT systems can also be categorized using traditional or modern technology.

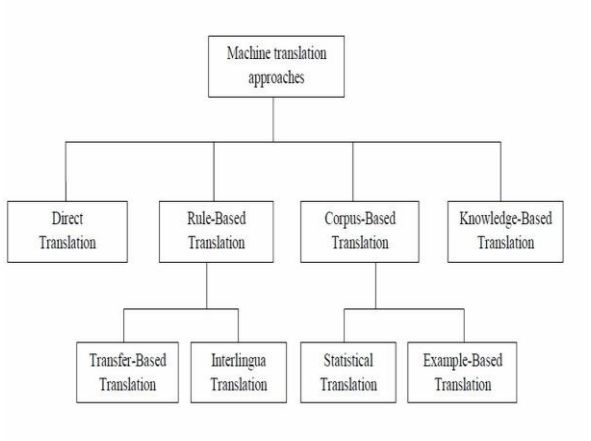


Figure 1: Machine Translation approaches

A. Direct MT

The direct MT form of MT is the most basic. It translates each word in a sentence from one language to another using a two-way dictionary. Uses very simple grammar rules. These plans are based on the principle that as the MT system should do as little work as possible. Specific MT systems take a monolithic approach to development, that is, they look at all the details of a single language.

Direct MT has the following characteristics:

- Little analysis of source language
- No parsing
- Reliance on large two-way dictionary

The standard procedure for specific translation programs can be summarized. The specific MT program begins with behavioural analysis. Morphological analysis removes the rise of morphological word-formation from the words to get the root word from the source language words. The next step in a straightforward MT program is to look up a bilingual dictionary. A bilingual dictionary is consulted to find the words of the target language that match the words of the source language. The final step in the MT-specific system is to re-syntactic reset. In syntactic rearrangement, word order is converted into something very similar to word order in the target language.

Direct Machine Translation works well with languages which have the same default sentence structure.

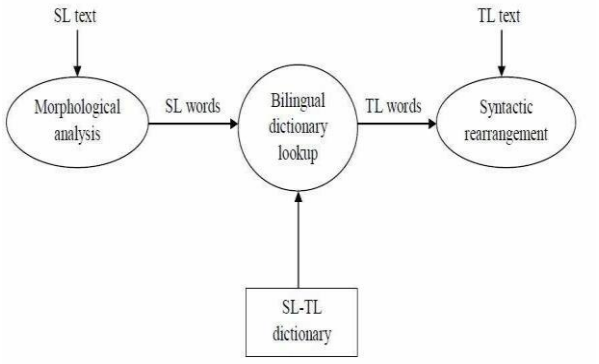


Figure 2: Direct Machine Translation

Advantages of Direct MT

Direct MT systems have the following advantages.

- The Translation is usually understood by the reader with little effort.

Disadvantage of Direct MT

The Direct MT systems have the following disadvantages.

- Direct MT includes character analysis only. It does not look at the structure and relationship between words.

- Direct MT programs are bilingual and cannot be translated into two different languages.
- Direct MT programs can be very expensive, in multilingual situations.
- Some source information may be lost in translation.

B. Rule-based MT

In rules-based programs, the source text is split and intermediate presentation is produced. Directed text writing is made from central representation.

These systems rely on the definition of morphology rules, syntax, lexical selection and transfer, semantic analysis and production.

• MT-based transfers

In this translation program, a database of translation rules is used to translate text from source to another language. Whenever a sentence is in line with one of the rules, or examples, it is translated directly using a dictionary. From source language to morphological and syntactic analysis to produce a certain type of Interlingual in the basic forms of the source language, from this it translates it into the target language types and from there a better translation is made to create the final step in the file.
The steps that are taken are shown in Figure.3

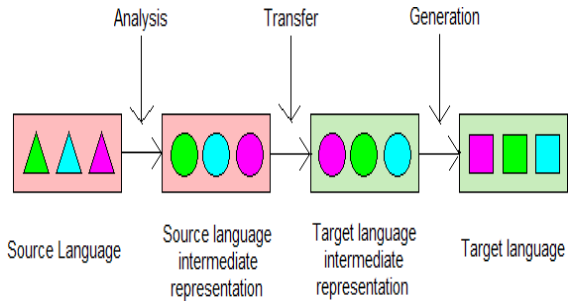


Figure 3: Description of transfer-based Machine Translation

Advantages of Transfer-Based MT
The transfer-based approach has the following advantages.

- It has a modular structure.
- The system easily handles ambiguities that transfer from one language to another.

Disadvantage of Transfer-Based MT
MT-based transmission systems have the following disadvantages.

- Some source information may be lost in translation.
- Bilingual Machine Inter Inter version is a sub-version of Direct Machine Translation.

The Interlingual Machine

Translation translates words into a general language designed for MT by simply translating them into more than one language. Figure 3 how different languages A, B, C, D can be translated into this program.
Advantages of Interlingual Machine Translation
Bilingual MT systems have lower advantages.

- Provides an objective-based presentation and can be used for programs such as retrieval of information.
- The multilingual system must resolve all ambiguities so that translation into any language can take place from bilingual representation.
- The system works best when several languages need to be translated because it only needs to be translated from the source language.
- In some areas, the bilingual approach may be used effectively.

Disadvantage of Interlingual Machine Translation
Interlingual MT systems have the following problems.

- The timing of this system is subject to the Direct Machine Translation program.
- The main problem lies in defining unintelligible representations (translated into two languages) that retain the meaning of the sentence.
- Defining the vocabulary of multilingualism is very difficult as different languages think of the world in different ways.
- There may be many ideas in a language or culture that are not represented in another language.

C. Corpus-Based Machine Translation

This is considered a new method of machine translation time. System-based systems are categorized by mathematical and model machine translation.

- Statistical Machine Translation (SMT)

A common idea in the SMT system is that translation will come from a word that can be translated less. The program consists of three different models. The Language Model (LM) lists the possible 'T' target language as P (T) probability. The Translation Model (TM), helps to calculate the probable conditions of the directed sentences given in the source sentence, P (T | S). Decoder enhances product potential for LM and TM.

II. STATISTICAL MACHINE TRANSLATION

The SMT system is based on the idea that every sentence in a language has a possible translation into another language. A sentence can be translated from one language to another in many ways. Mathematical translation methods are of the view that all sentences in the target language are a possible translation of input sentences. Figure 4 provides a framework for the Statistical Machine Translation program.

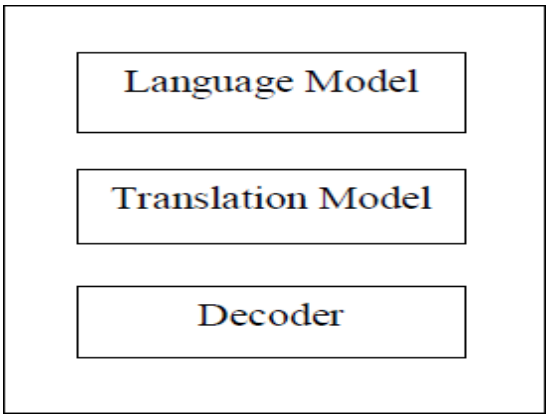


Figure 4: Statistical Machine Translation outline

A. *Language Model*

The language model provides the opportunity for a sentence. Chances are calculated using the n-gram model. The Model Model can be considered as a combination of single word opportunities given to all the preceding words in a sentence.

The goal of Statistical Machine Translation is to measure the probability (s) of a sentence. The sentence is separated from the product of the conditional probability. By using the chain rule, this is made possible. The probability of sentence
P (S), is broken down as the probability of individual words P(w).
P(s) = P(w 1, w 2, w 3 ,. , w n)
=P(w 1) P(w 2 |w 1) P(w 3, |w 1 w 2) P (w 4 |w 1 w 2 w 3)...P(w n |w 1 w 2. w n-1))

To calculate sentence possibilities, it is necessary to calculate the probability of a word, given the sequence of the preceding word. The n-gram model simplifies the task by speculating on the possibility of a given name for all previous words.
N-gram size 1 is called unigram; size 2 is a bigram (or, rarely, drawing); size 3 is the trigram; size 4 is a gram with size 5 or more just called n-gram.
Consider the following training data:

He was a strong King.
King ruled most parts of the world.
Probabilities for bigram model are as shown below:
P(there/<s>) = 0.67 P(was/there) = 0.4 P(king/a) = 1.0
P(a/<s>)=0.30 .... (2.2)
P(was/he) = 1.0 P(a/was) = 0.5 P(strong/a) = 0.2
P(king/strong)=0.23 .... (2.3)
P(ruled/he) = 1.0 P(most/rules) = 1.0 P(the/of) = 1.0 ... (2.4)
P(world/the) =0.30 P(ruled|king)=0.30 .... (2.5)
The probability of a sentence: 'A strong king ruled the world', can be computed as follows:
P(a/<s>)* P(strong/a)* P(king|strong)*P(ruled|king)*P(the/ruled)*P(world|the)
=0.30*0.2*0.23*0.30*0.28*.0.30
=0.00071

B. *Translation Model*

The Translation Model helps to calculate the probability of conditional P (T | S). Trained from the parallel corpus of the target-source pair. Since there is not a corpus large enough to allow for computer translation opportunities at the sentence level, so the process is divided into smaller units, e.g. Names or phrases and their possibilities have been studied [14]. The targeted translation of the source sentence is thought to be done from a literal source. For example, using a notation (T / S) to represent the S input sentence and its T translation. Using this text, the sentence is translated

(kutta bageche mae sooya | dog slept in the garden) (कुत्ता बगीचे में सोया था| dog slept in the garden).

C. *Decoder*

This section of SMT increases the chances of having the text translated. Selected words have the same hood as the translated translation.

Sentence T searches are performed to increase P (S | T) i.e.

$$Pr\ (S, T) = argmax\ P\ (T)\ P\ (S|T)$$

Here is the problem with the endless space to search. The use of featured search is suggested, where we keep a list of partial alignment. Search begins with a null hypothesis, meaning that a target sentence is found in the order of the source words we do not know. Represents this login sequence as (कुत्ता बगीचे में सोया | *),where * placeholder for the unknown sequence of source names. As the search progresses, add entries by adding one or more words to its concept. For example, extend the first entry for one or more of the following:

(कुत्ता बगीचे में सोया | dog slept (2))

Search ends when there is a perfect match in the list that promises more than any incomplete match.

III. PROBLEM FORMULATION

With each passing day the earth becomes a global valley. There are hundreds of languages spoken worldwide. The official languages of different regions and nations also differ according to their cultural and geographical diversity.

A. *Gap Analysis*

Most of the content available in digital format is in English. Content shown in English should be presented in a language that is not understood by the intended audience. There is a large proportion of people at the national and state level who do not understand English. It has created a language barrier in the digital age. Mission Translation (MT), can overcome this obstacle. In this article, the Statistical Based Machine Translation program for translating English text into Hindi has been suggested. English is the source language and Hindi is the target language.

B. *Objectives*

The objectives of the concept are below:
Understanding Language Model (LM), Translation Model (TM), and SMT Decoding sections.

- Creating Hindi LM using SRI's LM language model.
- Creating a TM model using GIZA ++ software.
- Making Hindi sentences using Moses software.
- System monitoring and evaluation.

C. *Methodology*

English to Hindi translation is done using Statistical Machine Translation (SMT). The main purpose of this program is to translate with less human effort. There are many tools related to LM, TM, SMT decoder. SMT has three main components of the program, Language Model, Model Model and search (decoder). LM encapsulates existing opportunities with regard to target language. TM encapsulates the possibilities for the inclusion of the target language name in the source language. For the development of 23LM, SRI international's SRILM Language Model toolkit is used. GIZA ++ is used for the construction of the Translation Model. In the code setting phase, Moses' software was used. The program is based on the Linux operating system. It will receive an English sentence from the forum and produce a release in Hindi.

IV. TOOLS USED FOR IMPLEMENTATION

There are many LM tools available. They are discussed as follows.

A. *The CMU Statistical Language Modelling(SLM) Toolkit*

Carnegie Mellon University (CMU) Statistical Language Modeling Toolkit is a collection of UNIX software tools designed to perform language modeling work for research purposes. Written by Roni Rosenfeld, and released in 1994.

B. *SRILM*

SRILM is a toolkit for creating and using Mathematical Language Models (Lms) developed by the SRI Speech Technology and Research Laboratory. It has been under construction since 1995. SRILM is free to download.

V. TRANSLATION MODEL TOOLS

There are many TM tools available for SMT programs. They are discussed as follows.

A. *GIZA++*

GIZA ++ is a tool developed by Franz Josef Och. Also the GIZA extension created by the Statistical Machine Translation team during a summer conference in 1999 at the Language and Speech Recognition Center at John-Hopkins University. This tool uses a variety of variables such as HMM and performs word alignment. GIZA ++ is available for free download.

B. *MGIZA*

MGIZA ++ is a multi-thread alignment tool based on GIZA ++. Expands GIZA ++ in many ways. It gives the impression of combining multiple fibers, and making good memory. It can also start training from any category, and continue training from any category. MGIZA is free to download.

VI. DECODER TOOLS

There are many different tools for the SMT system configuration phase. They are discussed as follows.

A. *Moses*

Moses is a Statistical Machine Translation program developed by Hieu Hoang and Philipp Koehn at the University of Edinburgh that allows automatic training of translation models of any language type. All that is needed is a compilation of translated parallel corpus. Moses works with SRILM to do the Language Model, and GIZA ++ to do the Translation Model. Moses is free.

B. *ISI ReWrite Decoder*

ISI ReWrite Decoder is software used to perform word processing (search) in the development of Statistical Machine Translation systems. Works with the CMU Tool-Statistical Language Modeling and GIZA ++ toolkit to perform translations from Source Language to Target Language.

VII. MERITS OF PROPOSED SYSTEM

A. *Faster Translation*

Using a machine translation system enables you to save your time while translating large texts.

B. *Low price*

When a professional translator translates your text, you have to pay enough for each page but we usually need a point of view, a common sense. In this case the machine translation system is reliable and works for us.

C. *Confidentiality*

Many people use machine translation systems to translate their private emails, because no one would agree to hand over their personal documents to a translator they don't know, or no one would be able to give other people financial documents.

D. *Universality*

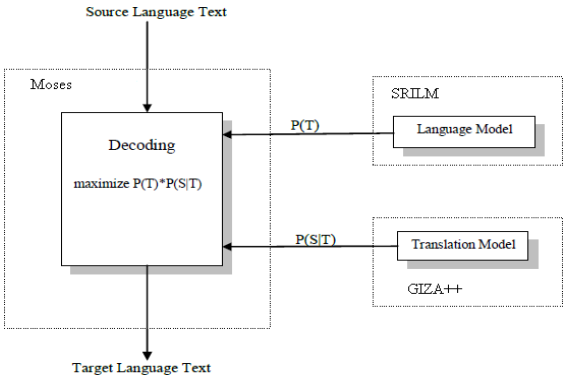
Usually a skilled translator becomes an expert in a specific field, but a typewriter can translate any text anywhere. The translation of special words should open the corresponding settings.

E. Online translation and translation of web page content

The beauty of online translation services is obvious. Online translation services are close at hand and you can quickly translate information about this service. Additionally you can translate any web page content with a search engine query using machine translation systems.

VIII. ARCHITECTURE DIAGRAM FOR PROPOSED METHOD

In this section, the implementation and implementation of the plan is discussed. These include corpus building, data editing, Model development, Model Model and decoder training.



A. Development of Corpus

The Mathematical Translation System uses the corresponding parallels between the two source and target languages. This compatible corpus is a prerequisite for prior training in Statistical Machine Translation. The proposed system used the same corporate English and Hindi sentences. A concise copy of more than 5000 sentences has been compiled containing the minimum sentences and life history of the freedom fighters based on their track record in the courts.

B. Architecture of English to Hindi Statistical Machine Translation System

The structures play an important role in building the SMT system. Language Model (LM), Translation Model (TM), decoder is used to make SMT. The language model is adapted from the target language. Decoder offers the possibility of a direct sentence given the source sentences.

C. Translation Model

The Translation Model (TM) considers the probability that the source sentence is 'S', with the targeted sentence given 'T'. Statistically, the TM counting opportunities are given as, P (S | T). Translation can be done based on a word or phrase. TM output is installed in Moses decoder. GIZA ++ and mkcls are used to develop the Translation model, which was developed.

D. Decoder

The decoder increases the chances of a synthetic sentence. It uses the function of argmax () to increase the chances. Moses' software freely available under open source licenses is used for decoder. Moses is compatible with SRILM and GIZA ++. The Moses decoder accepts as the inclusion of the text of the source language and creates the text of the target language. Opportunity files are accepted from TM and LM. The decoder can be set to interactive mode to perform the translation.

IX. IMPLEMENTATION AND DESCRIPTION

Details adjustment includes token insertion, cleaning, disposal of the corpse. Before undertaking a training program the details should be considered in advance. The issues that need to be addressed in the same corporations are as follows:

- Setting the default environment LC_ALL to C in Linux environment.
- Software requires one sentence per line. So there should be no empty lines in the corpus.
- Deleted sentences with more than 40 words. Sentence sentences from 1-40 have not been deleted.
- All parallel corpus sentences need to be included below. The top sentences should be changed to the lower case

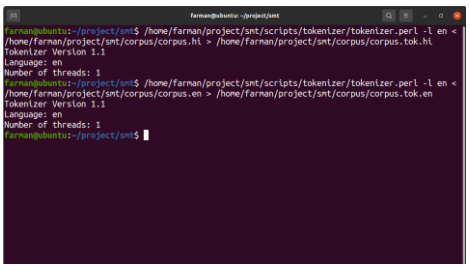
To prepare the data, used in the proposed system, PERL documents were used.

A. Tokenizing the corpus

Making corpus tokens uses Perl script. The input in this text is in corpus green and the output is a corpus with a token. Script created as provided in

$ /home/farman/project/smt/scripts/tokenizer/tokenizer.perl -l en < /home/farman/project/smt/corpus/corpus.hi > /home/farman/project/smt/corpus/corpus.tok.hi

$ /home/farman/project/smt/scripts/tokenizer/tokenizer.perl -l en < /home/farman/project/smt/corpus/corpus.en > /home/farman/project/smt/corpus/corpus.tok.en



Filtering out long sentences

Sorting long sentences using the PERL script, clean-corpus-n.perl. The release of tokenizer.perl was accepted as a clean-corpus-n-perl installation. This text removes long sentences from the corpus. It also removes unwanted space characters and blank lines. Long sentences, those that exceed the 40 word word limit. The system does not accept blank lines, which is why they are removed. GIZA ++ takes too long to train in long sentences. Clean-corpus-n.perl is used to reduce sentence lengths.

$/home/farman/project/smt/scripts/training/clean-corpus-n.perl /home/farman/project/smt/corpus/corpus.tok.en en /home/farman/project/smt/corpus/corpus.clean 1 40

B. Lowercaseing data

The data which is fed in for training the Moses software must be in small case. This is accomplished using *lowercase.perl*

/home/farman/project/smt/scripts/training /lowercase.perl < /home/farman/project/smt/corpus/corpus.tok.clean.en |more

X. LANGUAGE MODEL

For Language Model (LM), SR International's LM model (SRILM) is used. SRILM is compatible with Moses decoder and GIZA++ Translation Model (TM).

Command for generating 3-n-gram language model

/home/farman/project/smt/scripts/training/ngram-count –order 3 –text /home/farman/project/smt/corpus/corpus.tok.hi –lm hindi.lm –write count.cnt

Install Moses and respective packages

$mkdir project
$cd project
$sudo apt-get install build-essential git-core pkg-config automake libtool wget zlib1g-dev libbz2-dev

```
farman@ubuntu:~/project$ sudo apt-get install build-essential git-core pkg-con
Reading package lists... Done
Building dependency tree
Reading state information... Done
Note, selecting 'git' instead of 'git-core'
Note, selecting 'python-dev-is-python2' instead of 'python-dev'
pkg-config is already the newest version (0.29.1-0ubuntu4).
pkg-config set to manually installed.
wget is already the newest version (1.20.3-1ubuntu1).
wget set to manually installed.
The following additional packages will be installed:
  autoconf autotools-dev binutils binutils-common
  binutils-x86-64-linux-gnu bzip2-doc dpkg-dev fakeroot g++ g++-9 gcc
  gcc-9 git-man libalgorithm-diff-perl libalgorithm-diff-xs-perl
  libalgorithm-merge-perl libasan5 libatomic1 libbinutils libc-dev-bin
  libc6-dev libcrypt-dev libctf-nobfd0 libctf0 liberror-perl
  libexpat1-dev libfakeroot libgcc-9-dev libitm1 liblsan0 libltdl-dev
  libpython2-dev libpython2-stdlib libpython2.7 libpython2.7-dev
  libpython2.7-minimal libpython2.7-stdlib libquadmath0 libsigsegv2
  libstdc++-9-dev libtsan0 libubsan1 linux-libc-dev m4 make
  manpages-dev python-is-python2 python2 python2-dev python2-minimal
  python2.7 python2.7-dev python2.7-minimal
Suggested packages:
  autoconf-archive gnu-standards autoconf-doc gettext binutils-doc
  debian-keyring g++-multilib g++-9-multilib gcc-9-doc gcc-multilib
  flex bison gcc-doc gcc-9-multilib gcc-9-locales git-daemon-run
```

$sudo apt-get install libsoap-lite-perl

```
farman@ubuntu:~/project$ sudo apt-get install libsoap-lite-perl
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following additional packages will be installed:
  libclass-inspector-perl libconvert-binhex-perl libfcgi-perl libio-sessiondata-perl libmime-tools-perl libossp-uuid-perl libossp-uuid16
  libtask-weaken-perl libxmlrpc-lite-perl
Suggested packages:
  uuid libapache2-mod-perl2 libmime-lite-perl libnet-jabber-perl
The following NEW packages will be installed:
  libclass-inspector-perl libconvert-binhex-perl libfcgi-perl libio-sessiondata-perl libmime-tools-perl libossp-uuid-perl libossp-uuid16
  libsoap-lite-perl libtask-weaken-perl libxmlrpc-lite-perl
0 upgraded, 10 newly installed, 0 to remove and 0 not upgraded.
Need to get 591 kB of archives.
After this operation, 1,778 kB of additional disk space will be used.
Do you want to continue? [Y/n] y
Get:1 http://us.archive.ubuntu.com/ubuntu focal/universe amd64 libclass-inspector-perl all 1.36-1 [16.3 kB]
Get:2 http://us.archive.ubuntu.com/ubuntu focal/main amd64 libconvert-binhex-perl all 1.125-1 [29.7 kB]
Get:3 http://us.archive.ubuntu.com/ubuntu focal/main amd64 libfcgi-perl amd64 0.79-1 [33.1 kB]
Get:4 http://us.archive.ubuntu.com/ubuntu focal/universe amd64 libio-sessiondata-perl all 1.03-1 [5,606 B]
Get:5 http://us.archive.ubuntu.com/ubuntu focal/main amd64 libmime-tools-perl all 5.509-1 [192 kB]
Get:6 http://us.archive.ubuntu.com/ubuntu focal/universe amd64 libossp-uuid16 amd64 1.6.2-1.5build7 [28.4 kB]
Get:7 http://us.archive.ubuntu.com/ubuntu focal/universe amd64 libossp-uuid-perl amd64 1.6.2-1.5build7 [18.6 kB]
Get:8 http://us.archive.ubuntu.com/ubuntu focal/universe amd64 libtask-weaken-perl all 1.06-1 [8,700 B]
Get:9 http://us.archive.ubuntu.com/ubuntu focal/universe amd64 libsoap-lite-perl all 1.27-1 [236 kB]
Get:10 http://us.archive.ubuntu.com/ubuntu focal/universe amd64 libxmlrpc-lite-perl all 0.717-4 [22.5 kB]
```

$git clone https://github.com/moses-smt/mosesdecoder.git

```
farman@ubuntu:~/project$ git clone https://github.com/moses-smt/mosesdecoder.git
Cloning into 'mosesdecoder'...
remote: Enumerating objects: 498, done.
remote: Counting objects: 100% (498/498), done.
remote: Compressing objects: 100% (206/206), done.
remote: Total 148070 (delta 315), reused 433 (delta 289), pack-reused 147572
Receiving objects: 100% (148070/148070), 129.86 MiB | 1.06 MiB/s, done.
Resolving deltas: 100% (114341/114341), done.
farman@ubuntu:~/project$
```

$git clone https://github.com/moses-smt/giza-pp.git

```
farman@ubuntu:~/project$ git clone https://github.com/moses-smt/giza-pp.git
Cloning into 'giza-pp'...
remote: Enumerating objects: 301, done.
remote: Total 301 (delta 0), reused 0 (delta 0), pack-reused 301
Receiving objects: 100% (301/301), 259.42 KiB | 1.24 MiB/s, done.
Resolving deltas: 100% (207/207), done.
```

$cd giza-pp
$make

```
farman@ubuntu:~/project$ ls
giza-pp  mosesdecoder
farman@ubuntu:~/project$ cd giza-pp/
farman@ubuntu:~/project/giza-pp$ ls
GIZA++-v2  Makefile  mkcls-v2  README
farman@ubuntu:~/project/giza-pp$ make
make -C GIZA++-v2
make[1]: Entering directory '/home/farman/project/giza-pp/GIZA++-v2'
mkdir optimized/
g++    -Wall -Wno-parentheses -O3 -funroll-loops -DNDEBUG -DWORDINDEX_WITH_4_BYTE -DBINARY
/Parameter.o
Parameter.cpp: In function 'bool writeParameters(std::ofstream&, const ParSet&, int)':
Parameter.cpp:48:14: warning: ignoring return value of 'char* getcwd(char*, size_t)', dec
-result]
   48 |         getcwd(path,1024);
      |         ~~~~~~^~~~~~~~~~~
g++    -Wall -Wno-parentheses -O3 -funroll-loops -DNDEBUG -DWORDINDEX_WITH_4_BYTE -DBINARY
myassert.o
g++    -Wall -Wno-parentheses -O3 -funroll-loops -DNDEBUG -DWORDINDEX_WITH_4_BYTE -DBINARY
d/Perplexity.o
g++    -Wall -Wno-parentheses -O3 -funroll-loops -DNDEBUG -DWORDINDEX_WITH_4_BYTE -DBINARY
del1.o
```

$cd ..
$cd mosesdecoder
$mkdir tools

```
farman@ubuntu:~/project$ ls
giza-pp  mosesdecoder
farman@ubuntu:~/project$ cd mosesdecoder/
farman@ubuntu:~/project/mosesdecoder$ mkdir tools
farman@ubuntu:~/project/mosesdecoder$
```

Note: copying GIZA++ Binaries to mosesdeocder folder
$cp ../giza-pp/GIZA++-v2/GIZA++ ../giza-pp/GIZA++-v2/snt2cooc.out ../giza-pp/mkcls-v2/mkcls tools

downloading IRSTLM Toolkit from https://sourceforge.net/projects/irstlm/postdownload
$tar -xzf irstlm-5.80.08.tgz
$ cd irstlm-5.80.08/
$ ls

$ cd trunk

```
farman@ubuntu:~/project/mosesdecoder$ cd ..
farman@ubuntu:~/project$ ls
giza-pp  irstlm-5.80.08.tgz  mosesdecoder
farman@ubuntu:~/project$ tar -xzf irstlm-5.80.08.tgz
farman@ubuntu:~/project$ cd irstlm-5.80.08/
farman@ubuntu:~/project/irstlm-5.80.08$ ls
branches  regression-testing  trunk
farman@ubuntu:~/project/irstlm-5.80.08$ cd trunk
farman@ubuntu:~/project/irstlm-5.80.08/trunk$ ls
cmake           doc               make-release-dev         scripts
CMakeLists.txt  irstlm.xcodeproj  README                   src
configure.ac    Makefile.am       regenerate-makefiles.sh
Copyright       make-release      RELEASE
farman@ubuntu:~/project/irstlm-5.80.08/trunk$
```

$ cd ../..
$ mkdir irstlm
$ cd irstlm-5.80.08/
$ ./regenerate-makefiles.sh

```
farman@ubuntu: ~/project/irstlm-5.8
farman@ubuntu:~/project/irstlm-5.80.08/trunk$ cd ..
farman@ubuntu:~/project/irstlm-5.80.08$ cd ..
farman@ubuntu:~/project$ mkdir irstlm
farman@ubuntu:~/project$ cd irstlm-5.80.08/
farman@ubuntu:~/project/irstlm-5.80.08/trunk$ ./regenerate-makefiles.sh
Calling /usr/bin/autoreconf
configure.ac:10: warning: LT_INIT was called before AM_PROG_AR
/usr/share/aclocal-1.16/ar-lib.m4:13: AM_PROG_AR is expanded from...
configure.ac:10: the top level
configure.ac:10: warning: AC_PROG_LIBTOOL was called before AM_PROG_AR
/usr/share/aclocal-1.16/ar-lib.m4:13: AM_PROG_AR is expanded from...
configure.ac:10: the top level
configure.ac:10: warning: LT_INIT was called before AM_PROG_AR
/usr/share/aclocal-1.16/ar-lib.m4:13: AM_PROG_AR is expanded from...
configure.ac:10: the top level
configure.ac:10: warning: AC_PROG_LIBTOOL was called before AM_PROG_AR
/usr/share/aclocal-1.16/ar-lib.m4:13: AM_PROG_AR is expanded from...
configure.ac:10: the top level
configure.ac:10: warning: LT_INIT was called before AM_PROG_AR
aclocal.m4:9124: AM_PROG_AR is expanded from...
configure.ac:10: the top level
configure.ac:10: warning: AC_PROG_LIBTOOL was called before AM_PROG_AR
```

$ ./configure –prefix=/root/smt/irstlm

```
farman@ubuntu: ~/project/irstlm-
farman@ubuntu:~/project/irstlm-5.80.08/trunk$ ./configure -prefix=/root/smt/
checking for a BSD-compatible install... /usr/bin/install -c
checking whether build environment is sane... yes
checking for a thread-safe mkdir -p... /usr/bin/mkdir -p
checking for gawk... no
checking for mawk... mawk
checking whether make sets $(MAKE)... yes
checking whether make supports nested variables... yes
checking for gcc... gcc
checking whether the C compiler works... yes
checking for C compiler default output file name... a.out
checking for suffix of executables...
checking whether we are cross compiling... no
checking for suffix of object files... o
checking whether we are using the GNU C compiler... yes
checking whether gcc accepts -g... yes
checking for gcc option to accept ISO C89... none needed
checking whether gcc understands -c and -o together... yes
checking whether make supports the include directive... yes (GNU style)
checking dependency style of gcc... gcc3
checking for g++... g++
checking whether we are using the GNU C++ compiler... yes
```

$ make ---------- if error down grade gcc and g++ compiler

Following are command of down grade process

Append the following 2 lines at the end of the file /etc/apt/sources.list

deb http://dk.archive.ubuntu.com/ubuntu xenail main

deb http://dk.archive.ubuntu.com/ubuntu xenial universe

$ sudo apt update

$ sudo apt install gcc-4.9 and g++-4.9

$ gcc -v

$ sudo update-alternatives --remove-all gcc

$ sudo apt-get install gcc-4.8 g++-4.8

$ sudo update-alternatives --install /usr/bin/gcc gcc /usr/bin/gcc-4.8 50

$ sudo update-alternatives --install /usr/bin/g++ g++ /usr/bin/g++-4.8 50

$ sudo update-alternatives --config gcc

$ sudo update-alternatives --config g++

$ make install

Running Moses decoder

./moses -f ~/moses/mosesdecoder/trunk/scripts/training/moses-scripts/scripts/training/model/moses.ini

Output

Sr. No.	Input English Sentence	Output Hindi Sentence generated by the system
1	Students of the Dattatreya city Municipal corporation secondary school demonstrated their imagination power by creating the fictitious fort "Duttgarh".	महानगर पालिका अंतर्गत दत्तात्रय नगर माध्यमिक स्कूल के विद्यार्थियों ने काल्पनिक किला 'दत्तगढ़' बनाकर अपनी कल्पनाशक्ति का परिचय दिया।
2	With encouragement from Principal Sandhya Medpallivaar the teachers and students built the fort out of clay.	प्रधानाध्यापक संध्या मेडपल्लीवार के प्रोत्साहित करने पर शिक्षकों व विद्यार्थियों ने मिट्टी से किले का निर्माण किया।
3	Rajesh Gavre, the President of the MNPA teachers association, honoured the school by presenting the award.	मनपा शिक्षक संघ के अध्यक्ष राजेश गवरे ने स्कूल को भेंट देकर सराहना की।
4	Ramesh Saatpute examined the fort.	किले का परीक्षण रमेश सातपुते ने किया।
5	Students like Nikhil Kavle, Darshan Gedekar, Sahil Meshram participated in building the fort.	किला निर्माण में निखिल कावले, दर्शन गेड़ेकर, साहिल मेश्राम इन विद्यार्थियों ने सहभाग लिया।
6	Narender Barai, the President of the District Art Teachers' association, Shekhar Vanskar, a cashier, Ajay Gundamwar, a member of Gajanan Mehar provided guidance to the students.	जिला कलाध्यापक संघ के अध्यक्ष नरेंद्र बारई, कोषाध्यक्ष शेखर वानस्कर, सदस्य अजय गुंडमवार, गजानन मेहर ने विद्यार्थियों का मार्गदर्शन किया।
7	Nagarsevak, Reeta Mule presented messages from well-wishers.	नगरसेवक रीता मुले ने सदिच्छा भेंट दी।
8	Rohtak. Akhil Bhartiya Janwadi Mahila Samiti and DYFI jointly launched a state-wide campaign against corruption in jobs, fraud and hold-ups.	रोहतक. नौकरियों में भ्रष्टाचार, फर्जीवाड़े व लूट- खसोट के खिलाफ अखिल भारतीय जनवादी महिला समिति और डीवाईएफआई ने संयुक्त रूप से प्रदेशव्यापी अभियान शुरू किया है।
9	Through this state-wide signature campaign, 10 Lakh (1 million)signatures across the state will be collected and handed over to the Governor.	इस राज्यव्यापी हस्ताक्षर अभियान के माध्यम से प्रदेश भर में 10 लाख हस्ताक्षर करवाकर राज्यपाल को सौंपे जाएंगे।
10	The signature campaign started on Friday at the new bus stand.	शुक्रवार को नए बस स्टैंड पर हस्ताक्षर अभियान की शुरुआत की गई।
11	On this occasion, the SFI state secretary Manoj Kumar, Anju, District Secretary of the Committee, State joint-secretary of the Youth council Vinod Deshwal, Sumit, Anju, Rakesh Kumari, Geeta, Sonu, Rajesh Kumar, Sangeeta, Meena, Veena Malik, Sangeeta, Hawa Singh and Ajit were present.	इस मौके पर एसएफआई के राज्य सचिव मनोज कुमार, समिति की जिला सचिव अंजू, नौजवान सभा के प्रदेश सहसचिव विनोद देशवाल, सुमित, अंजू, राकेश कुमारी, गीता, सोनू, राजेश कुमार, संगीता, मीना, वीना मलिक, संगीता, हवा सिंह और अजीत मौजूद रहे।
12	A very sad incident occurred in Maloya village, which is located on the outskirts of the city, where a newly married woman committed suicide by hanging herself from the fan.	शहरी के छोर पर बसे मलोया गांव में शुक्रवार को हुई दुखद घटना में नवविवाहिता कोमल ने पंखे से फंदा लगाकर जान दे दी।
13	Komal had married only two months ago.	कोमल की शादी को अभी दो महीने ही हुए थे।

14	A police investigation found that Komal was worried about her financial situation and so she took this step.	पुलिस की प्रारंभिक जांच में सामने आया है कि कोमल आर्थिक तंगी से परेशान थी व इसी वजह से उसने यह कदम उठाया।
15	According to the description, the details of the death of Komal only came to light when one of Komal's cousin went to the third floor of their house to do some cleaning.	विवरण के अनुसार कोमल द्वारा मौत को गले लगाने के घटनाक्रम का पता शुक्रवार देर दोपहर को उस समय लगा जब उसकी चचेरी बहन तीसरी मंजिल की छत पर बने कमरे में सफाई करने गई।
16	Seeing the body hanging from the fan the girl screamed and immediately informed the in-laws about it.	शव को पंखे से लटकते देख वह चीख उठी और उसके सुसर को इस वाकये की सूचना दी।
17	Komal was immediately taken to the Multi-speciality Government Hospital, Sector-16, where she was declared dead on arrival.	कोमल को तत्काल गवर्नमेंट मल्टीस्पेशिएलिटी अस्पताल, सेक्टर-16 में ले जाया गया जहां डाक्टरों ने उसे मृत लाई गई घोषित कर दिया।
18	Komal's husband Kulvinder is unemployed.	कोमल का पति कुलविंदर कुछ कामकाज नहीं करता।
19	Before committing suicide, Komal wrote with henna on her left hand that she was committing suicide of her own free will as the difficulties in her life were too much for her.	कोमल ने यह कदम उठाने से पहले अपने बाएं हाथ पर मेंहदी से लिखा कि वह अपनी मर्जी से खुदकुशी कर रही है, मेरे जीवन की मुश्किलें ही मेरी मौत की वजह हैं।
20	According to the details received, Komal's father had passed away a few years ago, her mother is mentally ill and her brother is studying in a government school.	विवरण के अनुसार कोमल के पिता की कई साल पहले मृत्यु हो चुकी है उसकी मां, मानसिक रोगी है और एक भाई सरकारी स्कूल में पढ़ता है।
21	In the meantime, the police has kept the body in the mortuary of the hospital in Sector-16.	बहरहाल, पुलिस ने शव को सेक्टर-16 के अस्पताल की मोर्चरी में रखवा दिया है।
22	Only after a post-mortem will the police be able to find out the actual cause of death.	पोस्टमार्टम रिपोर्ट के बाद मृत्यु के वास्तविक कारणों का पता चल सकेगा।
23	Sri Lankan selectors have selected 16 members of the team for the up coming limited-over match series, which is going to take place from the 10th of November, against New Zealand.	श्रीलंकाई चयनकर्ताओं ने आगामी 10 नवंबर से न्यूजीलैंड के खिलाफ होने वाले सीमित ओवर के द्विपक्षीय सीरीज के लिए 16 सदस्यीय श्रीलंकाई टीम का चयन किया है।
24	The selectors have included only one new face, the 24 year old Ashan Priyanjana, while they also called Dimuth Karunarathne back to the team after two years.	चयनकर्ताओं ने टीम में एक मात्र नया चेहरा 24 वर्षीय हरफनमौला खिलाड़ी एशान प्रियंजना को शामिल किया है जबकि दिमुथ करुणारत्ने को दो वर्ष बाद फिर से टीम में जगह दी गई है।
25	The Sri Lankan team will play three ODI (One Day International) matches and two T-20 matches against the Kiwis from 10th to 21st November.	श्रीलंकाई टीम कीवियों के खिलाफ 10-21 नवंबर तक तीन एक दिवसीय जबकि दो टी-20 मैच खेलेगी।
26	The first and second one day matches will be played in Hambantota, while the third one day match will be played in Dambulla.	पहला और दूसरा एक दिवसीय मुकाबला हंबनटोटा में जबकि तीसरा वनडे मुकाबला दांबुला में खेला जाएगा।
27	Both T-20 matches will be played in Pallekele.	वहीं दोनों टी-20 मुकाबले पाल्लेकेल में खेले जाएंगे।

XI CONCLUSION

In this thesis, English to Hindi SMT system has been developed. The SMT is a part of corpus based MT system which requires parallel corpus before undertaking translation. A parallel corpus of 5000 English and Hindi sentences was used to train the system. The SMT system developed accepts English sentences as input and generates corresponding translation in Hindi. The translation of 90 sentences was evaluated using human evaluation method. On the parameters of fluency and adequacy a geometric average of 2.693 and 2.93 was calculated respectively.

The quality of the translated text are often depends upon the dimensions of the corpus and therefore the quality of the corpus.

There can be following future directions for English to Hindi SMT system.

The work can be extended to include multilingual corpus of different languages in the source-target pair. The target and source languages can be increased from present one language.

• The system can also be put in the web-based portal to translate content of one web page in English to Hindi.

• A mobile application can also be developed in which message containing English text is sent to the client in Hindi language.

• The corpus can be preprocessed to change its clause structure for improving the quality of translation.

• The translated text can be reordered and processed to overcome grammatical mistakes which will be part of post-processing. This will improve score of human evaluation

REFERENCES

[1] A. Stolcke, "SRILM-An Extensible Language Modelling Toolkit". Proc. Intl. Conf. on Spoken Language Processing, vol. 2, pp.901–904, Denver.

[2] "ngram-count",[Online].Available: http://www.speech.sri.com/projects/srilm/manpages/ngram-count.1. html/

[3] "Statistical Machine Translation System User Manual and Code Guide", [Online]. Available: http://www.statmt.org/moses/manual/manual.pdf/

[4] S. Charles and S. David, "Overview of Statistical Machine Translation", John Hopkins University, AMTA2006, [Online]. Available: http://www.cs.umass.edu/~dasmith/smt-tutorial.ppt

[5] F.J. Och., "GIZA++: Training of statistical translation models", [Online]. Available at: http://fjoch.com/GIZA++.html.

[6] S. Singh, M. Dalal, V. Vachhani, P.Bhattacha

[7] P. F. Brown, S. De. Pietra, V. D. Pietra and R. Mercer, "The mathematics of statistical machine translation: parameter estimation". "Journal Computational Linguistics", vol. 19, no.3, June 1993.

[8] Stolcke, "Guide on how-to install and build SRI LM", [Online] Available: http://www.speech.sri.com/projects/srilm/docs /INSTALL.

[9] Charniak and Eugene, "Introduction to artificial intelligence", Boston: Addison-Wesley, 1984.

[10] "Natural language processing", [Online].Available: http://en.wikipedia.org/wiki/ Natural_language_processing

[11] "Statistical Machine Translation", [Online]. Available: http://www.comp.nus.edu.sg/~huangyun/ ebook/ 2008_Statistical_Machine_Translation. pdf

[12] P. Kohen, "Moses: Open Source Toolkit for Statistical Machine Translation." Proceedings of the ACL 2007 Demo and Poster Sessions, pp. 177–180, Prague, June 2007.

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

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

[15] S.K. Dwivedi and P. P. Sukadeve, "Machine Translation System Indian Perspectives", roceeding of Journal of Computer Science Vol. 6 No. 10. Pp 1082-1087, May 2010.
[16] G. Athens, "Automated Translation of Indian languages", ACM News, Magazine Communications of the ACM, DOI: 10.1145.
[17] "Machine Translation", [Online]. Available: http://en.wikipedia.org/wiki/ Machine_translation
[18] "Rule-based machine translation", [Online]. Available: http://en.wikipedia.org/wiki/Rule-based_machine_translation
[19] "Transfer-based machine translation", [Online]. Available: http://en.wikipedia.org/wiki/Transfer-based_machine_translation
[20] "Interlingual machine translation",[Online]. Available: http://en.wikipedia.org/wiki/ Interlingual_machine_translation
[21] "Dictionary-based machine translation", [Online].Available: http://en.wikipedia.org/wiki/Dictionary-based_machine_translation
[22] "Example-based machine translation", [Online].Available: http://en.wikipedia.org/wiki/Example-based_machine_translation
[23] "Statistical machine translation", [Online]. Available, http://en.wikipedia.org/wiki/ Statistical_machine_translation
[24] Tanveer Siddiqui and U.S. Tiwary, "Natural language Processing and Information Retrieval", New Delhi, Oxford Press, 2008.
[25] "Machine Translation", [Online]. Available, http://faculty.ksu.edu.sa/homiedan/ Publications/ Machine%20Translation.pdf
[26] "Machine Translation ", [Online],Available: http://www.ida.liu.se/~729G11/projekt/ studentpapper-10/maria-hedblom.pdf
[27] "CMU toolkit manual", [Online], Available: http://mi.eng.cam.ac.uk/~prc14/ toolkit_documentation.html
[28] G. Singh and G. Singh Lehal," A Punjabi to Hindi Machine Translation System", Coling 2008: Companion volume- Posters and demonstrations, Manchester, August 2008.