ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

An Approach for Telugu to English Phrase Based Statistical Machine Translation System

Mr G. Lakshmikanth, M.Tech
Associate Professor,
Dept of CSE,
Sree Rama Engineering College, Tirupathi.

Smt.B.Dhana Lakshmi, M.Tech
Assistant Professor,
Dept of CSE,
Sree Rama Engineering College, Tirupathi.

Abstract:

Machine Translation is an important part of Natural Language Processing. It refers to using machine to convert one natural language to another. For example in Indian languages we have ethics like Ramayana, Mahabharata etc., which are life transforming stories, should be made available in all other languages. For this purpose, I had developed Telugu to English Machine Translation System. In this system Telugu sentences are given as input and we get out as English sentences.

In Telugu, the format of the sentence will be Subject-Object-Verb (SOV) type but in English we have Subject-Verb-Object (SVO) format. Statistical Machine Translation is a part of Machine Translation that strives to use machine learning paradigm towards translating text. Statistical Machine Translation consists of Language Model (LM), Translation Model (TM) and Decoder. In this thesis, Telugu to English Statistical Machine Translation system has been developed. The development of Language Model, Translation Model and decoder is done by making use of software's available in Linux environment.

IRST Language Model Tool kit (IRSTLM) for Language Model, GIZA++ and mkcls for Translation Model, Moses for decoding, has been used in this system. LM computes the probability of target language sentences by using large vocabulary of bilingual parallel corpus. TM calculates the probability of target sentences given the source sentence and the decoder maximizes the probability of translated text of target language. A parallel corpus of 10760 sentences in Telugu and English has been used in training of the system. The system was evaluated using manual evaluation method and a geometric average score of 8.693, 6.93 on the parameters of fluency and adequacy respectively, were found.

Keywords:

Statistical Machine Translation; Machine Learning; Bilingual parallel corpus.

I. Introduction:

The technology is reaching new heights, right from conception of ideas up to the practical implementation. It is important, that equal emphasis is put to remove the language divide which causes communication gap among different sections of societies. Natural Language Processing (NLP) is the field that strives to fill this gap. Machine Translation (MT) mainly deals with transformation of one language to another. Coming to the MT scenarios in India, it has enormous scope due to many regional languages of India. It is pertinent that majority of the population in India are fluent in regional languages such as Hindi, Telugu, Tamil, Punjabi etc. [9]. Given such a scenario, MT can be used to provide an interface of regional language.

1.1 Machine Translation:

Machine Translation (MT) is a sub-field of computational linguistics that investigates the use of computer software to translate text or speech from one natural language to another [24]. At its basic level, MT performs simple substitution of words in one natural language for words in another. Current machine translation software often allows for customization by domain or profession (such as weather reports), improving output by limiting the scope of allowable substitutions.

This technique is particularly effective in domains where formal or formulaic language is used. It follows that machine translation of government and legal documents more readily produces usable output than conversation or less standardized text [24].

1.1.1 Need for MT:

Machine Translation system are needed to translate literary works which from any language into native languages. The literary work is fed to the MT system and translation is done. Such MT systems can break the language barriers by making available work rich sources of literature available to people across the world.MT also overcomes the technological barriers. Most of the Indian languages ethics like Ramayana, Mahabharat etc., which are life transforming stories, should be made available to other languages. This has let to digital divide in which only small section of society can understand the content presented in digital format. MT can help in this regard to overcome the digital divide.

1.1.2 Problems in MT:

There are several structural and stylistic differences among languages, which make automatic translation a difficult task. Some of these issues are as follows.

Word order:

Word order in languages differs. Some classification can be done by naming the typical order of subject (S), verb (V) and object (O) in a sentence . Some languages have word orders as SOV. The target language may have a different word order. In such cases, word to word translation is difficult [15]. For example Telugu Language has SOV and English Language has SVO sentence structure.

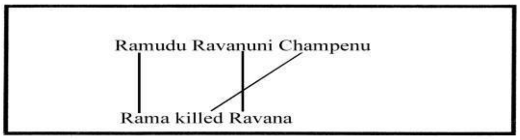


fig 1.1: Example of reordering

Word sense:

The same word may have different senses when being translated to another language. The selection of right word specific to the context is important [15].

Pronoun Resolution:

The problem of not resolving the pronominal references is important for machine translation. Unresolved references can lead to incorrect translation [15].

Idioms:

An idiomatic expression may convey a different meaning, that what is evident from its words. For example, an idiom in English language *'Jack of all trades'*, would not convey the intend meaning when translated into Telugu language [15].

Ambiguity:

In computational linguistics, Word Sense disambigua tion (WSD) is an open problem of natural language processing, which governs the process of identifying which sense of a word (i.e. meaning) is used in a sentence, when the word has multiple meanings [15].

1.1.3 Approaches to MT:

There are four approaches to Machine Translation. These are discussed as follows.

Rule-based MT:

A Rule-based MT system parses the source text and produces an intermediate representation, which may be a parse tree or some abstract representation [15].

Direct-based MT:

A direct-based MT system carries out word-by-word translation with the help of a bilingual dictionary, usually followed by some syntactic rearrangement [15].

Corpus-based MT:

Corpus based MT systems require sentence-aligned parallel text for each language pair. The corpus based approach is further classified into statistical and example based machine translation approaches [15].

Knowledge-based MT:

Early MT systems are characterized by the syntax. Semantic features are attached to the syntactic structures and semantic processing occurs only after syntactic processing. Semantic-based approaches to language analysis have been introduced by AI researchers.

The approached require a large knowledge-base that includes both ontological and lexical knowledge [15].

The Statistical Machine Translation (SMT) is part of corpus based Machine Translation. SMT requires less human effort to undertake translation. SMT is a machine translation paradigm where translations are generated on the basis of statistical models. These statistical models parameters are derived from the analysis of bilingual text corpora. In this thesis, proposed Telugu to English Statistical Machine Translation system has been presented. The translation system has three modules: Language Model, Translation model, and Decoder. For Language Model, SR International's Language Model toolkit (IRSTLM) is used to develop Language Model for English. English is the target language for the system. To develop Translation Model, open source GIZA++ software is used. To perform decoding, open source Moses software is used.

II. LITERATURE SURVEY:

In this section, Telugu to English SMT system literature review, for this work is discussed.

2.1 MT approaches:

MT systems can be classified according to the means by which they perform translation. A classification of MT approaches is given in Figure 2.1 [14]. The MT systems can also be classified according to use of traditional or modern technology.

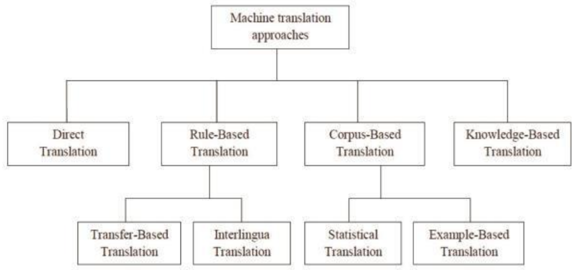


Figure 2.1: Machine Translation approaches [15]

2.1.1 Direct MT:

Direct MT form of MT is the most basic one. It translates the individual words in a sentence from one language to another using a two-way dictionary. It makes use of very simple grammar rules. These systems are based upon the principle that as MT system should do as little work as possible. Direct MT systems take a monolithic approach towards development, *i.e.*, they consider all the details of one language pair. Direct MT has following characteristics:

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

The general procedure for direct translation systems can be summarized as shown in Figure 2.2. The direct MT system starts with morphological analysis. Morphological analysis removes morphological inflections from the words to get the root word from the source language words. The next step in direct MT system is bilingual dictionary lookup. A bilingual dictionary is looked up to get the target-language words corresponding to the source-language words. The last step in direct MT system is syntactic rearrangement. In syntactic rearrangement, the word order is changed to that which best matches the word order of the target language [15].

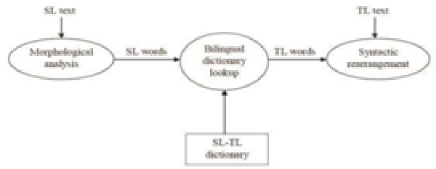


Figure 2.2: Direct Machine Translation [15]

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

2.1.2 Rule-based MT:

In rule-based systems, the source text is parsed and an intermediate representation is produced. The target language text is generated from the intermediate representation. These systems rely on the specification of rules for morphology, syntax, lexical selection and transfer, semantic analysis and generation [14].

Transfer based MT:

In this translation system, a database of translation rules is used to translate text from source to target language. Whenever a sentence matches one of the rules, or examples, it is translated directly using a dictionary. It goes from the source language to a morphological and syntactic analysis to produce a sort of Interlingua on the base forms of the source language, from this it translates it to the base forms of the target language and from there a better translation is made to create the final step in the translation. The steps which are performed are shown in Figure 2.3.

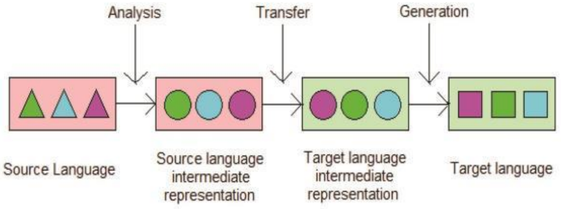

Analysis
Transfer
Generation
Source Language
Source language intermediate representation
Target language intermediate representation
Target language

Figure 2.3: Description of Transfer-Based Machine Translation [25]

The major modules in transfer based MT is as follows.

Analysis: Analysis phase is used to produce source language structure [15].Transfer: Transfer phase is used to transfer source language representation to a target level representation [15]. Generation: Generation phase is used to generate target language text using target level structure [15].

• Interlingua Machine Translation:

Inter is a sub version of Direct Machine Translation. The Interlingua Machine Translation converts words into a universal language that is created for the MT simply to translate it to more than one language. Figure 2.4 shows how different languages A, B, C, D can be translated through this system.

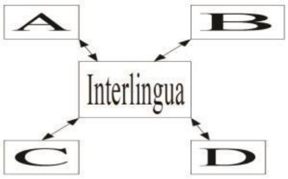


Figure 2.4: Interlingua language system [26] Advantages of Interlingua Machine Translation

Interlingua MT systems have below mentioned advantages. It gives a meaning-based representation and can be used in applications like information retrieval.

An Interlingua system has to resolve all the ambiguities so that translation to any language can take place from the Interlingua representation. The system is more practical when several languages are to be interpreted since it only needs to translate it from the source language. Figure 2.5 shows how language A can be translated into several languages. For specific domains, Interlingua approach can be used successfully .

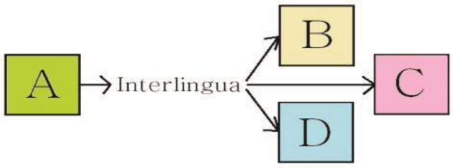


Figure 2.5: Multilingual MT system with Interlingua approach [25] Disadvantage of Interlingua Machine Translation

Interlingua MT systems have following disadvantages.

Time efficiency of this system is lower than the Direct Machine Translation system. Major problem lies in defining a universal abstract (Interlingua) representation which preserves the meaning of a sentence. Defining a vocabulary for a universal Interlingua is extremely difficult as different languages conceptualize the world in different ways. There may be many concepts in a language or culture which lack representation in another language [15].

2.1.3 Corpus Based Machine Translation:

This is considered as a new approach of the era for machine translation. The corpus based systems are classified into statistical and example-based Machine Translation.

Statistical Machine Translation (SMT):

The general idea in SMT system is that the translation will be from the most likely translated word. The system consists of three different models. The Language Model (LM) computes the probability of the target language *'T'* as probability *P(T)*. The Translation Model (TM), helps to compute the conditional probability of target sentences given the source sentence, *P(T|S)*. Decoder maximizes the product of LM and TM probabilities.

Example based Machine Translation:

Example based systems use previous translation examples to generate translations for an input provided. When an input sentence is presented to the system, it retrieves a similar source sentence from the example-base and its translation. The system then adapts the example translation to generate the translation of the input sentence.

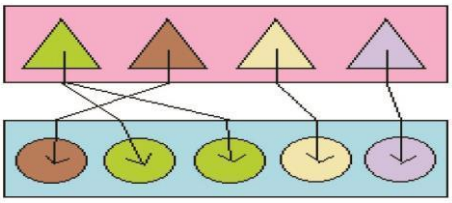


Figure 2.7: Translation Template of a phrase in two different languages [15]

Translation templates are a bilingual pair of sentences or phrases where words are coupled and replaced by variables. The goal is to have large corpus to be able to directly translate word after word in a sentence based on the *translation template* [15].Figure 2.7 shows the *translation templates* of a phrase in two different languages. Differences between Example-based MT and statistical-based MT systems are given in Table 2.1.

Table 2.1: Difference between example and statistical MT

Example-based MT	**Statistical-based MT**
Example-based MT systems use variety of linguistic resources such as dictionaries and thesauri, *etc.*, to translate text.	Statistical-based MT uses purely statistical based methods in aligning the words and generation of texts.

Number MT techniques are being combined to undertake translation. There is still a lot of research being done to couple a number of MT systems. There are some hybrid systems which are being proposed. One such system is Generation-Heavy MT (GHMT) [14].

2.1. 4 Phrase Based MT:

The sequence-based approach to Statistical Machine Translation is best known *as* phrase-based SMT (PB-SMT). "Phrase" is a consecutive sequence of words, not the linguistic phrase which means sequence of words that Junctions as a single unit in the language syntax. The strength of this approach lies in its ability to handle collocation relations between words within the sentence. This makes it compatible with the frequency-based nature of the language, also known as property of selection restriction. The phrase-based approach overcomes many of the weaknesses of the word-based IBM models. The ability to directly describe relation between strings of different lengths improves modeling considerably. This provides a direct relation between the source context and the target word selection. In addition, the phrases provide a much better model for local reordering than (he distortion probability. Finally, in translation, some content-weak words do not carry over to the other language. This means that some source words seem to be deleted in translation, and some target words seem to appear out of nowhere.The fundamental definition of Statistical Machine Translation is reformulated in connection with PB-SMT. Ochs & Ney propose a direct modeling of the translation process based on the maximum entropy approach. Instead of employing the Bayesian Theorem, the posterior probability $P\left(\frac{t_1^I}{s_1^J}\right)$ "expressed as a set of 'M' sub models $h_m(t_1^I, s_1^J)$.

Which are weighted by a model parameter λ_m:

$$P\left(\frac{t_1^I}{s_1^J}\right) = \frac{\exp\left(\sum_{m=1}^{M}\lambda_m h_m(t_1^I, s_1^J)\right)}{\sum_{t_1^I}\exp\sum_{m=1}^{M}\lambda_m h_m(t_1^I, s_1^J)}$$

Since the normalizing denominator only depends on the source sentence, it has no influence on the maximization of the entire probability. This leads to the following equations.

$$\hat{s}_1^I = \arg\max_{e_1^I}\left(\frac{t_1^I}{s_1^J}\right) = \arg\max_{e_1^I}\sum^{M}\lambda_m h_m(t_1^I, s_1^J)$$

$_I\ P\ |\qquad J\ \sum_m$

$1\ \left(^s1\right)\qquad 1\quad \bar{1}$

This model is also known as the log-linear model. The noisy channel approach can be viewed as a special case of this model where it contains two equally weighted sub models, namely the language and the translation model.

A major advantage of this generalization is that it provides a basis for easily enriching the translation process with additional information. It is very easy to add a new feature related to the translation process by just adding it to 'M^5 with appropriate weight.

2.1.5 Knowledge-based MT:

Early MT systems are characterized by the use of syntax. There was little semantic analysis. The synthesis and analysis is restricted in early MT systems to sentence level. Semantic-based approaches to language analysis have been introduced AI researchers. The approaches require a large knowledge –base that includes both ontological and lexical knowledge [15].

2.2 Statistical Machine Translation:

The SMT system is based on the view that every sentence in a language has a possible translation in another language. A sentence can be translated from one language to another in many possible ways. Statistical translation approaches take the view that every sentence in the target language is a possible translation of the input sentences [14]. Figure 2.8 gives the outline of Statistical Machine Translation system.

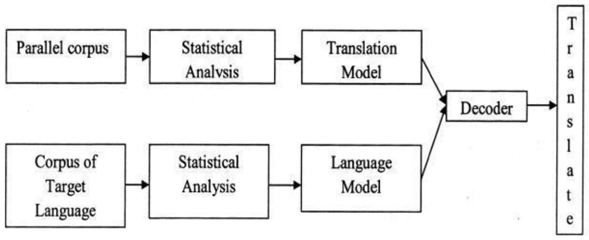


Figure 2.8: Outline Statistical Machine Translation system

2.2.1 Language Model:

A language model gives the probability of a sentence. The probability is computed using *n-gram* model. Language Model can be considered as computation of the probability of single word given all of the words that precede it in a sentence [25]. The goal of Statistical Machine Translation is to estimate the probability (likelihood). of a sentence. A sentence is decomposed into the product of conditional probability. By using chain rule, this is made possible as shown in 2.1. The probability of sentence P (S), is broken down as the probability of individual words P(w).

$P(s) = P(w_1, w_2, w3,....., wn)$
$=P(w_1)\quad P(w_2|w_1)\quad P(w_{3,}|w_1w_2)\quad P(w_4|w_1w_2w_3)\ldots P(w_n|w_1w_2\ldots w_{n-1}))$ … (2.1)

In order to calculate sentence probability, it is required to calculate the probability of

a word, given the sequence of word preceding it. An *n-gram* model simplifies the task

by approximating the probability of a word given all the previous words.

An *n*-gram of size 1 is referred to as a *unigram*; size 2 is a *bigram* (or, less commonly,

a *digram*); size 3 is a *trigram*; size 4 is a *four-gram* and size 5 or more is simply called a *n-gram*. Consider the following training set of data given in Figure 2.9:

There was a King

He was a strong King.

King ruled most parts of the world.

Figure 2.9: Training set of data for LM

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 ... (2.6)

2.2.2 Translation Model:

The Translation Model helps to compute the conditional probability *P(T|S)*. It is trained from parallel corpus of target-source pairs.

As no corpus is large enough to allow the computation translation model probabilities at sentence level, so the process is broken down into smaller units, *e.g.*, words or phrases and their probabilities learnt [14]. The target translation of source sentence is thought of as being generated from source word by word. For example, using the notation *(T/S)* to represent an input sentence *S* and its translation *T*.
Using this notation, sentence is translated as given in

(dog slept in the garden | kukka thotalo nidrapoye) (dog slept in the garden| కుక్క తోటలో నిద్రపోయ్) One

possible alignment for the pair of sentences can be represented as given in (dog slept in the garden | కుక్క(1), నిద్రపోయ్(4), లో (3), తోట(2))A number of

alignments are possible. For simplicity, word by word alignment of translation model is considered. The above set of alignment is denoted as *A(S, T)*. If length of target is *l* and that of source is *m* than there are *lm* different alignments are possible and all connection for each target position are equally likely, therefore order of words in T and S does not affect *P (T|S)* and likelihood of *(T|S)* can be defined in terms of the conditional probability *P (T, a/S)* as shown in 2.9: P(S|T) = sum P(S, a/T) ... (2.9) The sum is over the elements of alignment set, A(S, T). English word has only exactly one connection. For the alignment, P(dog slept in the garden| కుక్క తోటలో నిద్రపోయ్), can

be computed by multiplying the translation Probabilities T(dog| కుక్క(1)), T(garden | తోట (6)),

T (in | లో (4)), T (null | the(5)), T (slept | నిద్రపోయ్

(2))

To generate target sentence from source sentence, we have to follow the steps as given below: i). Select the length of *S* with probability *L* where L=P[length (S)=*m*] is a constant *i.e.* All lengths are assumed to be equally likely with probability *L*. ii). Select an alignment with probability *P(a|S)*. There are (l+1)m possible alignments [23]. Assuming all possible alignments are equally likely, the probability of alignment a, *P(a|S)*, is as shown in 2.10 P(a|S) = L x1/(l+1)m ... (2.10) iii). Select the *j*th English word with a probability The joint likelihood of Telugu string and an alignment given an English string is given in 2.11 P(S, a/T) = P(a/T)*P (S/a, T) ... 2.11) *T* is the probability of seeing *Sj* in source sentence, given *Taj* in target sentence. The alignment is determined by specifying the values of *aj* for *j* from 1 to *m*, each of which can take value from 0 to l.

2.2.3 Decoder:

This phase of SMT maximizes the probability of translated text. The words are chosen which have maximum like hood of being the translated translation [7]. Search for sentence T is performed that maximizes *P (S|T) i.e.*

Pr (S, T) = argmax P (T) P (S|T) ... (2.12)
Here problem is the infinite space to be searched. The use of stacked search is suggested, in which we maintain a list of partial alignment hypothesis [7]. Search starts with null hypothesis, which means that the target sentence is obtained from a sequence of source words that we do not know. We represent this entry sequence as (కుక్క తోటలో నిద్రపోయ్ |*), where

* is a place holder for an unknown sequence of source words. As the search proceeds, it extends entries in the list by adding one or more additional words to its hypothesis. For example, extend initial entry to one or more of the following entries:

(కుక్క తోటలో నిద్రపోయ్ | dog slept (2))

The search terminates when there is a complete alignment in the list that is more promising than any of the incomplete alignments.

III. EXPIREMENTAL FRAMEWORK:

In this section, the design and implementation of the system has been discussed. This includes development of corpus, data preparation, development of Language Model, Translation Model and training of decoder.

3.1 Development of Corpus:

Statistical Machine Translation system makes use of a parallel corpus of source and target language pairs. This parallel corpus is necessary requirement before undertaking training in Statistical Machine Translation. The proposed system has used parallel corpus of Telugu and English sentences. A parallel corpus of more than 43,500 sentences has been developed from which consist of small sentences and the life history of freedom fighters with reference to their trail in courts.

3.1.1 Architecture of Telugu to English Statistical Machine Translation System

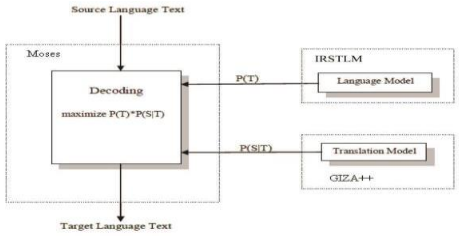


Figure 3.1: Architecture of Statistical Machine Translation system

The architecture forms the central role in making up SMT system. Language Model (LM), Translation Model (TM), decoder are used in undertaking SMT. Language Model is prepared from the target language. Decoder gives the probability of target sentence given the source sentences. The architecture of the system is shown in Figure 3.1.

3.1.2 Language Model:

Language Model (LM) gives the probability of a sentence. The probability of a sentence depends upon the probability of individual words. n-gram is a sequence of words [17]. LM is developed for the target language. If 'T' is the target language, LM computes 'P(|T)' and feed this input to the decoder software.IRST International's Language Model (IRSTLM) for LM is used. IRSTLM is available freely for research purposes from their website http://www.speech.sri.com/projects/IRSTLM/download.html.

3.1.3 Translation Model:

The Translation Model (TM) computes the probability of source sentence 'S', for a given target sentence 'T'. Mathematically, the probability being computed by TM is given as, P(S|T). Translations can be done word based or phrase based [5]. The output of TM is fed into Moses decoder. GIZA++ along with mkcls is used to develop Translation model, which is developed.

3.1.4 Decoder:

The decoder maximizes the probability of the generated sentence. It makes use of the argmax () function to maximize the probability. *Moses* software which is freely available under open source licenses is used for decoder. *Moses* is compatible with IRSTLM and *GIZA++*. *Moses* decoder accepts as input the source language text and generates the target language text. The probability files are accepted from TM and LM. The decoder can be set in interactive mode to for doing translation [12].

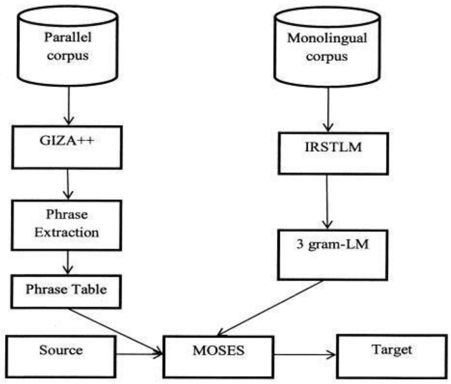


Figure 3.2: Architecture of Telugu to English Translation System

3.2 Preparation of Data:

Preparation of data involves tokenizing, cleaning, lowercasing the corpus. Before undertaking the training of the system the data must be pre-processed. The issues which need to be addressed in parallel corpus are as follows: To set the environment variable LC_ALL to C in Linux environment. The software needs one sentence per line. So there should be no empty lines in the corpus. The sentences having word limit more than 40 words are removed. The sentences having word limit from 1-40 are not removed.All sentences of parallel corpus need to be in lowercased. The uppercased sentences need to be changed to lower case [4].For the preparation of data, used in proposed system, PERL scripts have been used.

3.3 Language Model:

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

Ngram-count:

Ngram-count counts the number of *n-gram* of the corpus.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Ngram-count also builds the language model from the generated counts [2]. The format of LM is also shown by ngram-format file. The command for generating language model is given in 4.5. ./ngram-count –order 3 –text corpus_new4.lowercased.hi –lm hindi.lm –write count.cnt …(4.5)The keyword \data\ indicates the beginning of lm file. The total count of individual ngrams, found in the corpus is then mentioned after \data\ keyword. For each n-gram (1-gram, 2-gram, etc.), there are individual sub-sections. Each sub-section starts with conditional probability of the n-gram. This probability is to the base of log 10. This is followed by the word which constitutes n-gram [2].

3.4 Translation Model:

The software that aids in developing Translation Model is GIZA++. GIZA++ is extension of GIZA software. GIZA++ includes a lot of additional features [5]. The extensions of GIZA++ were designed and written by Franz Josef Och.

GIZA++ has following features:
Implements full IBM-4 alignment model
Implements IBM-5: dependency on word classes, smoothing, Implements HMM alignment model Smoothing for fertility, distortion/alignment parameters Improved perplexity calculation for models IBM-1, IBM-2 and HMM [5].The latest version of *Moses* software embeds calls to *GIZA++* and *mkcls* software's, hence there no need to call them separately.

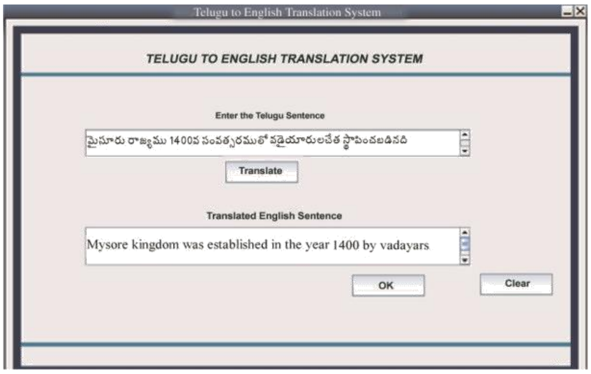
3.5 Decoder:

Moses software helps in decoding stage of SMT. It allows us to train translation models for any language pair. The pre-requisite for the translation is already translated, parallel corpus.

3.5.1 Training Moses decoder:

Moses toolkit embeds calls to Translation Model (GIZA++) software inside its training script. As a result, the phrase and reordering table get created. The script that does this is called train-factored-model.perl. Consider a Telugu sentence:

మైసూర్ రాజ్యం వడ్డెయరులుచేత ద్వారా సంవత్సరం 1400 లో స్థాపించబడింది decoder accepted this input in the interactive mode. The result of this translation is shown in Figure 3.8.


Telugu to English Translation System
TELUGU TO ENGLISH TRANSLATION SYSTEM
Enter the Telugu Sentence
Translate
Translated English Sentence
OK
Clear

మైసూరు రాజ్యము 1400వ సంవత్సరములో వడైయారులచేత స్థాపించబడినది

Mysore kingdom was established in the year 1400 by vadayars

Figure 3.8 Result of English sentence

మైసూర్ రాజ్యం వడ్డెయరులుచేత ద్వారా సంవత్సరం 1400 లో స్థాపించబడింది.

By executing Moses in interactive mode, 100 entences were translated to English language. Table 4.8 gives the English sentences along with the corresponding translation done by Moses into English language.

IV. RESULTS AND EVALUATION:

The proposed Telugu to English SMT system, accepted Telugu language sentences as input and gave English sentences as output. The translation of 100 Telugu sentences was done into English language. There are two ways of evaluating any MT system.Evaluation can be done automatically or manually. In this thesis, manual evaluation method has been used. The translation was evaluated on the parameters of fluency and adequacy.

Adequacy is defined as the degree to which the reference sentence is conveyed in the translation. Fluency refers to the grammatical accuracy of the translated text. Adequacy and Fluency for a given translation has different levels on which it can be evaluated [8]. The levels on which fluency and adequacy were evaluated are given in Table 4.1 and 4.2.

Table 4.1: Levels of fluency

Parameter	Definition	Ranking
Perfect	This indicates good grammar of sentence that is translated	4
Fair	The translated sentence is easy to understand but has lack correct grammar	3
Acceptable	The translated sentence is broken, but is understandable with efforts	2
Nonsense	The translated sentence is not clear	1

Table 4.2: Levels of adequacy

Parameter	Definition	Ranking
All	The translated sentence express all the meaning of the source sentence	4
Most	The translated sentence conveys all the meaning of source sentence	3
Some	The translated sentence conveys some meaning	2
None	The translated sentence does has no meaning conveyed	1

The above mentioned parameters (fluency and adequacy) were evaluated by three persons. The geometric average of the parameters was taken and the results are shown in Table 4.3.The geometric average of the individual parameters, (fluency and adequacy) was taken and the scores were as shown in Table 4.3.

Table 4.3: Result of SMT evaluation

	Fluency	Adequacy
Geometric average	2.693	2.93

V. CONCLUSION AND FUTURE WORK:

5.1 Conclusion:

In this thesis, Telugu to English 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 10760 Telugu and English sentences was used to train the system. The SMT system developed accepts Telugu sentences as input and generates corresponding translation in English. The translation of 100 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 can be depends upon the size of the corpus and the quality of the corpus.

5.2 Future Scope:

. There can be following future directions for Telugu to English 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 Telugu to English.

. A mobile application can also be developed in which message containing Telugu text is sent to the client in English 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, “IRSTLM-An Extensible Language Modeling Toolkit”. Proc. ntl. Conf. on Spoken Language Processing, vol. 2, pp.901–904, Denver.

[2] “ngram-count”, [Online]. Available: ttp://www.speech.sri.com/projects/IRSTLM/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,

[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. Bhattacharyya, and O. P. Damani, “Hindi generation from Interlingua (UNL)”,[Online] Available: http://www.cse.iitb.ac.in/~damani/papers/MTSummit0.pdf

[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] A. Stolcke, “Guide on how-to install and build SRI LM”, [Online] Available: http://www.speech.sri.com/projects/IRSTLM/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”, Proceeding 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

[19] “Interlingual machine translation”,[Online].

[20] “Dictionary-based machine translation”, [Online].Available: http://en.wikipedia.org/wiki/Dictionary-based_machine_translation

[21] “Example-based machine translation”, [Online].Available: http://en.wikipedia.org/wiki/Example-based_machine_translation

[22] “Statistical machine translation”, [Online]. Available, http://en.wikipedia.org/wiki/Statistical_machine_translation

[23] Tanveer Siddiqui and U.S. Tiwary, “Natural language Processing and Information Retrieval”, New Delhi, Oxford Press, 2008.

[24] “Machine Translation”, [Online]. Available, http://faculty.ksu.edu.sa/homiedan/Publications/Machine%20Translation.pdf