IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 37, No. 5, October, 2023, pp. 1109-1120
Journal homepage: http://iieta.org/journals/ria

Lexical Based Reordering Models for English to Telugu Machine Translation

Bandi Vamsi[1], Ali Al Bataineh[2*], Bhanu Prakash Doppala[3]

[1] Department of Artificial Intelligence, Madanapalle Institute of Technology & Science, Madanapalle 517325, Andhra Pradesh, India
[2] Artificial Intelligence Center, Norwich University, Northfield 05663, United States
[3] Data Analytics, Generation Australia, 88 Phillip St, Sydney 2000, Australia

Corresponding Author Email: aalbatai@norwich.edu

https://doi.org/10.18280/ria.370503

Received: 11 August 2023
Revised: 5 October 2023
Accepted: 11 October 2023
Available online: 31 October 2023

Keywords:
machine translation, phrase translation, reordering, orientation, parallel corpus

ABSTRACT

Telugu is one of the commonly spoken regional language in India. It is mostly spoken among the states of Andhra Pradesh and Telangana. In rural areas it is difficult for the people to understand the non-regional language specially while at the time of government works, land dealing transactions etc. Due to this there is a scope to develop a machine translation model from English to Telugu. The machine translation is an automatic technique of translating one language to another through Language Processing approach. To understand the Telugu language translation, the structural comparisons are done among English and Telugu languages to attain standard outcome. In this work, Lexical based reordering statistical model (LBRSM) is used for language conversion. This analyzes the language structure outcomes between word, phrase and hierarchical based models for the translation quality purpose. To maintain good quality translation TER and BLEU metrics 62.01 and 29.07 are considered for finding n-gram exact matches. From this work, the Phrase based reordering statistical model (PBRSM) achieved better results when compared with other system models in both training and testing phases.

1. INTRODUCTION

According to language census rate of India in 2011, Telugu is considered as the 4th most spoken language among 9,500 regional languages [1, 2]. As per the ‘Ethnologue’ report in terms of population, Telugu is the 13th language spoken around the world [3]. Machine translation (MT) itself defines the translation of text from source language to target language. It is a subset of Artificial Intelligence (AI) technology. This is used to understand various languages while communicating with others [4]. It helps in dividing the barriers of a language and makes it easier in international cooperation for a common man to understand. The translation approach is a complex task because handling human language need verification at different stages [5]. At every particular stage this need to be analysed which becomes more complex due to large size of finite words. The main aim of MT is to achieve meaningful sentence from one language to another language translation by using reordering mechanisms [6].

In rural areas, people usually speak in their own regional language. While in case of remote areas the slang of the people varies depending on their region or place. The public servants may come from different localities and thus in such scenario public face difficulty in understanding their language. In such situations, a general translation is required to understand the basic structure of the language to communicate with them. Hence, MT provides an opportunity to overcome this difficulty.

The Machine Translation (MT) can be achieved by rule-based, dictionary-based and corpus-based approaches. The rule-based is an intermediate translation approach in which all the rules are designed manually by language experts [7]. The dictionary-based approach relies on dictionary entries and word by word translations. These approaches require complete language grammatical structure data for translation from both ends which also consumes more time on complex datasets [8]. Due to this, the corpus-based approach is reliable in automatic translation process. This approach contains pre-translated sequence of phrases instead of words. The MT uses corpora statistical data as a main vocabulary re-ordering source for word, phrase and hierarchical models. The terms reordering refers in rearranging the words in the given input sentence that which is to be translated into the target language. In this work, complete evaluation of phrase based reordering statistical model (PBRSM) is used for translation when compared with other models named as Word based reordering statistical model (WBRSM) and Hierarchical based reordering statistical model (HBRSM). PBRSM uses word by word independent translation derived from simple words. This independent usage of translation requires larger datasets making it more complex. Hence, PBRSM is evolved by arranging these words into phrases for translation which becomes easier.

In this work, we used various statistical reordering models such as: Phrase based reordering statistical model (PBRSM), Word based reordering statistical model (WBRSM) and Hierarchical based reordering statistical model (HBRSM) for translation of English to Telugu language. The objectives that are used in this work as follows:

a) Initially the complete sentence is fragmented into phrases.
b) The reordering of words are then identified by using alignment matrix for translation purpose.

c) The heuristic translation table is constructed for mapping of words among English to Telugu language and vice versa.
d) The alignment points are identified based on union/intersection through phrase table mapping.
e) To attain the standards of target language different reordering mechanisms are used for n-gram models.

The remaining contents of this work are organized as follows. Section-2 briefs the Related Work, Section-3 deals with Methodology of the study, Section-4 discusses about Results of the lexical models and Section-5 elaborates Conclusion and Future Scope.

2. RELATED WORK

This section covers all of the associated work that has been done by other researchers in a comparable field of study.

Dungarwal et al. [9] proposed that heuristic strategies outperform generative features in the activity of pair of phrase retrieval. The use of quantity, instance, and tree bases neighboring details as components that aids in the translation of English to Hindi. By demonstrating a translation model enhancement using pre- and post-processing elements. Authors preorder the input text to correspond to the specific language in order to resolve the structural differentiation among English and Hindi.

Many to one and one to one alignment interconnections are used in the GIZA++ development of IBM frameworks. Because of this limitation, several linguistic sentences cannot be properly aligned. To efficiently overcome these limitations, Och and Ney [10, 11] suggested symmetric function of heuristic strategies. Various methods for integrating word mappings in order to symmetrize directed quantitative orientation models.

Utilizing synthetic dataset, Dutta Chowdhury et al. [12] evaluated the influence of training a multi neural learning based model with feature representation for a limited language pair images for Hindi and English translation. Based on a conventional English image corpus, the authors created a synthesized training set as well as an individually compiled expansion dataset for Hindi.

Reddy and Hanumanthappa [13] Machine translation, which effectively converts data from one to another natural language by maintaining its meaning is the main goal of NLP. In order to identify and translate English into Kannada or Telugu, this research provides a unique model for MT that uses rule and dictionary based approaches.

A dynamic rule-based MT model from English to Telugu was proposed by Lingam et al. [14]. To accomplish this, a collection of classification rules, a training and testing sets of English and Telugu phrases, and an English to Telugu vocabulary have been constructed. The fundamental challenge with MT is how to handle parts of speech. The selection of the proper postposition in Telugu depends on a number of factors, including age, time, place, content, and others. Depending on the situation, the right preposition should be chosen from those that have varying semantics. Words and phrases that are often used can be properly processed and converted using this approach.

Discourse translation requires extra care because the statements must be translated while maintaining the context in sight. It is best to read the first and second phrases together rather than separately. The subject of the current study is the translation of two categories of compound, discourse and difficult words in a sentence English into Telugu. This work is about creating algorithms that translate complicated and elaborate phrases from English to Suryakanthi and Sharma [15].

Neural MT (NMT) is able to translate English into Indian languages, particularly Telugu. In addition to NMT, an attempt is required to increase accuracy by using a strong preprocessing approach. Appropriate preprocessing will play a smaller but still significant effect in enhancing accuracy. NMT needs a huge volume of parallel corpus to do the translations. Telugu and English are resource-constrained languages, creating a parallel corpus is expensive [16].

The structural abnormalities among Hindi and English create a challenge to provide accurate translations. Translation is done in this work using the phrase based MT model. The three major functional elements of this model are translation, sequencing, and training set. This study analyses the effects of different configurations of these characteristics on the accuracy of automatic English to Hindi translations [17].

According the previous works [9-17], the translation among one language to other languages are chosen based on leading languages in top countries using MT. These translations are very effective in terms of standards and quality. But in cases of regional and rural areas the frequency of understanding level is low. The language structure of leading language is simple by which it is easy to achieve high standard translation quality among source and target languages. However, for the regional languages the structure is complex and require various reordering mechanisms to maintain the good translation quality. This translation quality depends on maintaining the standards in grammar and syntax of the resulting language. To overcome this limitation, in this work we used various lexical based statistical model types like word, phrase and hierarchical n-grams models for translation of English to Telugu language. The summary of these previous works is shown in Table 1.

3. METHODOLOGY

The methodology of this work focuses on various technical aspects such as word alignment, reordering mechanisms, n-gram findings etc. The word alignment is a major aspect in MT for identifying word-to-word references in a pair of sentences. The fundamental task of this alignment is to detect the relation among the words of sentence given as input in various languages. On the other hand, reordering also equally plays a key role in rearranging the sentence in target language. It maintains efficiency and quality of the sentence.

This section deals with dividing a sentence into group of phrases under Section 3.1., language translation based on word alignment matrix is given in Section 3.2., heuristic based approaches using union or intersection operations are discussed in Section 3.3., various re-ordering mechanism required for language translation is given under Section 3.4., and the parts of the lexical model is discussed in Section 3.5. All these sections show about the working procedure of LBRSM in detail. The development of reordering models is discussed under Section 3.6. The summary of the dataset used in this study is given in Section 3.7

Table 1. Summary of existing works used in this study

Author	Model	Approach	Result	Advantages
Dungarwal et al. 2004 [9]	SMT	Phrase based translation between Hindi and English languages	BLEU: 27.62	This approach mainly focuses on phrase based only without n-grams comparison.
Och and Ney 2003 [10]	Statistical alignment model	German to English language Translation	Precision outcomes for Intersection: 91.5 and Union: 63.4	Heuristic approaches are not sufficient to translate complex sentences. Re-ordering mechanisms are required.
Och and Ney. 2004 [11]	SMT	Phrase based translation between Chinese to English language	BLEU: 53.0 for 2-gram BLEU: 55.2 for 3-gram	Word and phrase translations carried out on single words.
Dutta Chowdhury et al. 2018 [12]	Phrase and Neural based MT	English to Hindi language translation	BLEU for phrase based: 21.6 BLEU for NMT: 23.3	Parallel corpus data is available in image format which requires more processing time for training. During training, maintaining all character fonts with respect to stokes is crucial.
Reddy and Hanumanthappa 2013 [13]	Rule based MT	English to Kannada / Telugu language translation	No coverage	Defining and handling a new rule for language translation becomes difficult if the length of grams in a sentence increase.
Lingam et al. 2014 [14]	Rule based MT	English to Telugu language translation	Testing accuracy: 92%	Telugu language has complex structure. To handle this type of language, re-ordering is required instead of simple translation rules.
Suryakanthi and Sharma 2015 [15]	MT	English to Telugu language translation	For compound and complexed perfect sentences, the translation scores are 50% and 10%.	Handling complex and compound sentences requires re-ordering mechanisms. Without this the translation quality becomes poor.
Raju et al. 2021 [16]	Corpus based neural MT	English to Telugu language translation	BLEU: 47.70	Handling unknown words requires huge size of training data.
Babhulgaonkar and Sonavane 2022 [17]	Phrase based MT	English to Hindi language translation	BLEU: 23.71 TER: 67.21 For 5-gram	All types of re-ordering orientation mechanisms are used only in phrase based. Comparative analysis is required among all types of translation models.
Current model	Lexical based reordering statistical model (LBRSM)	Translation between English and Telugu languages using phrase, word and hierarchical models	BLEU: 29.87 TER: 55.6 for 6-gram	This model performs translation between source and target language based on alignment matrix, heuristic approach, re-ordering mechanism like distance and lexical based with all types of orientations.

3.1 Formal framework

When a sentence is given as input, it is divided into a group of phrases and direct mapping of phrases is made.

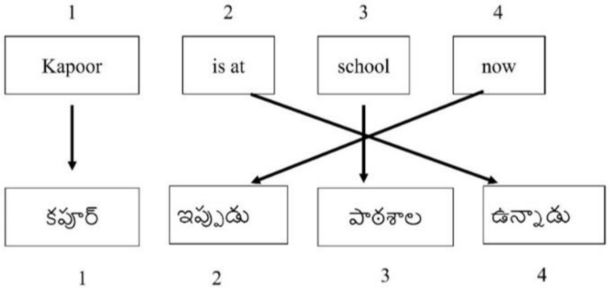


Figure 1. Dividing the sentence into phrases

Figure 1 describes partitioning of a sentence into sequence of words named as phrases. The probability score is calculated by PBRSM model which is interlinked with pairs of phrases. These scores are stored in pre-defined phrase table. In this work, different parameters of PBRSM model are analyzed to evaluate the effects on quality of sentence translation from English to Telugu language. This is done through a three-step process such as training, translation and testing. In training phase, phrase alignment is identified by giving parallel corpus English as source and Telugu as target language. Then in translation phase, given source sentence is divided randomly into phrases and by using translation table the final target language is obtained. In the final testing phase, the quality of the sentence is identified by reordering the target phrase to get meaningful Telugu sentence.

According to Naïve Bayes theorem [9], deriving the probability of translating English sentence 'e' into Telugu language 't' is formulated in Eq. (1).

$$p(e|t) = argmax_e \, p(t|e) \, p(e) \tag{1}$$

While at the time of decoding, the Telugu sentence 't' is divided into group of 'T' phrases 't_j^T'. We consider an equal distribution of probabilities among every feasible division. Every Telugu phrase 't_j' in 't_j^T' is transformed to English phrase 'e_i'. A probability distribution is given by '$\varphi(t_j|e_i)$' to represent phrase translation.

An estimate of the absolute distortion distribution of probability '$d(start_i, end_{i-1})$' is used to describe the rearrangement of the resulting English phrase. Here, '$start_i$' represent the initial alignment of Telugu sentence which is to be translated into 'i^{th}' phrase of English sentence. The resulting position of Telugu sentence is given by 'end_{i-1}' which is converted into English sentence of $(i-1)^{th}$ phrase.

To estimate the length of target language, we assume a word cost factor 'ω' for every converted English phrase along with 'p_{LM}' which is an n-gram model language. This can be used to improve the model performance. Generally, to attain distortion for longer outcomes, this value is set to more than 1.

In brief, the accurate English target sentence 'e_{best}' represents a Telugu source sentence 't' in accordance with the proposed model is given by Eq. (2).

$$e_{best} = argmax_e \ \varphi(t_j|e_i) \\ d(start_i, end_{i-1}) \ p_{LM} \ \omega^{length(e)} \tag{2}$$

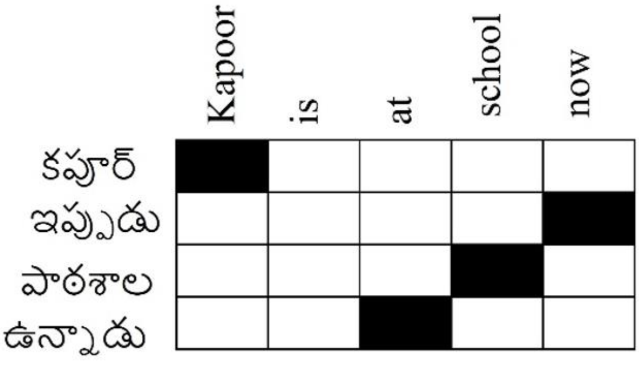
The initial translation is carried by '$\varphi(t_j|e_i)$', reordering of phrases is given by '$d(start_i, end_{i-1})$' and meaningful and grammatical order of target English phrase is given by '$p_{LM} \ \omega^{length(e)}$'.

3.2 Alignment matrix for translation

The main fundamental approach in MT model is alignment of word. This alignment among the words can be identified by word alignment matrix as shown in Figure 2. The size of this matrix is dynamic and the size varies depending up on the input sentence. The black cells in the matrix represents the alignment point among English sentence 'Kapoor is at school now' and Telugu sentence 'కపూర్ ఇప్పుడు పాటశాలలో ఉన్నాడు'. In the sentence there may or may not be alignment points within words. This occurs because of few words in English do not have corresponding clear translation in Telugu language. From Figure 2 it is observed that, the English word 'is' is not aligned in Telugu language.

During the translation phase, it is complex to get word to word mapping from source to target language. It becomes very difficult in case of idiomatic sentences due to new and missing words obtained in input language. In this phase, the alignment of words is complex while reordering, deletions and insertions. If $e_i^E = e_1, e_2, e_3 \dots . e_E$ is a source sentence in English and $t_j^T = t_1, t_2, t_3 \dots . t_T$ is a target sentence in Telugu, which should be aligned by words. The subset 'M' is a word alignment of the Cartesian product according to the places of words available in input and output sentence is given by Eq. (3).

$$M \subseteq \{ (i,j) \ where \ i = 1,2, \dots . E \ and \ j = 1,2, \dots T\} \tag{3}$$



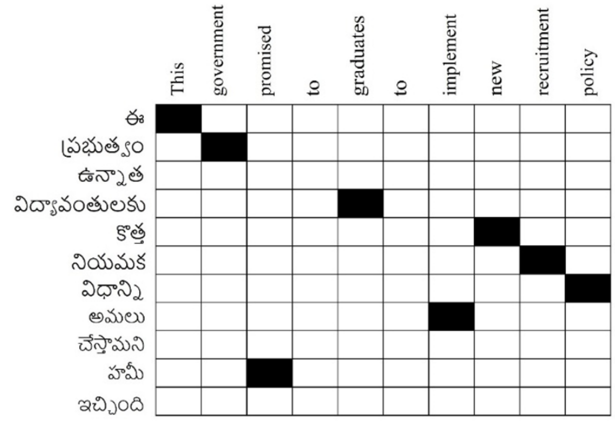
	Kapoor	is	at	school	now
కపూర్	■				
ఇప్పుడు					■
పాఠశాల				■	
ఉన్నాడు			■		

Figure 2. Word alignment matrix

3.3 Heuristic based phrase translation table

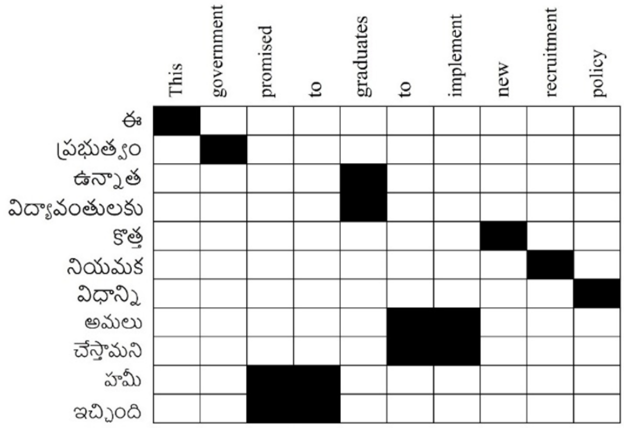
The words are positioned in both the ways of conversion process, like from English language to Telugu language and vice-versa. Once after the alignment, different word arrangement heuristics are involved to produce symmetric sequence through the initial two arrangements. The word positioning heuristics originates by the convergence of word positions gathered through this. For enhancing the positions, these are inspected from other places in the combination of these initial arrangements. The convergence and combination of English to Telugu and from Telugu to English word arrangement for a converted sentence is represented in Figure 3 and Figure 4.

The highlighted region or cells in phrase translation table indicates the matched words between English to Telugu language. For better understanding, these matched cells are shown in 'black' color as represented in Figure 3. By converting a sentence from Telugu to English language requires more combination of highlighted cells as shown in Figure 4. The empty cells in both examples represents no mapping between source to target language translation.



	This	government	promised	to	graduates	to	implement	new	recruitment	policy
ఈ	■									
ప్రభుత్వం		■								
ఉన్నాత										
విద్యావంతులకు					■					
కొత్త								■		
నియమక									■	
విధాన్ని										■
అమలు							■			
చేస్తామని										
హామీ			■							
ఇచ్చింది										

Figure 3. Phrase table for English to Telugu sentence mapping



	This	government	promised	to	graduates	to	implement	new	recruitment	policy
ఈ	■									
ప్రభుత్వం		■								
ఉన్నాత					■					
విద్యావంతులకు					■					
కొత్త								■		
నియమక									■	
విధాన్ని										■
అమలు						■	■			
చేస్తామని						■	■			
హామీ			■	■						
ఇచ్చింది			■	■						

Figure 4. Phrase table for Telugu to English sentence mapping

The heuristic alignment of phrase begins with intersection operation of the word orientations produced by these two fundamental approaches such as mappings from English to Telugu and Telugu to English. Then these enhanced connections are aligned by union operation. Figure 5 depicts the union or intersection of aligned points translated from Telugu to English and English to Telugu.

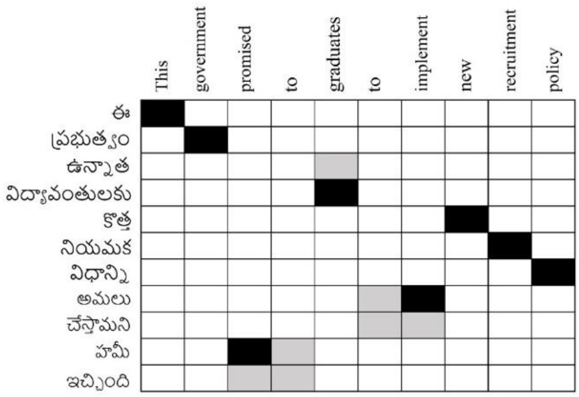


Figure 5. Aligned points from union/intersection operations

In this work, according to Och and Ney [10, 11] there are eight number of conversions which converts any function into symmetric method in terms of variables. These are helpful in analyzing their effect on English to Telugu conversion standard. Let 'X_1' define translation of source to target language and 'X_2' define translation of target to source language from available training data. The eight heuristic conversions are as follows:

1. **union:** All aligned points are considered by performing union operation on X_1 and X_2 which is represented as $X_1 \cup X_2$.
2. **intersection:** Taking aligned points which are common among X_1 and X_2 and is denoted as $X_1 \cap X_2$.
3. **srctotgt:** Indicate word alignment obtained from English to Telugu language and is referred as X_1.
4. **tgttosrc:** Indicate word alignment obtained from Telugu to English language and is referred as X_2.
5. **grow:** The term itself refer gathering the nearby aligned points along with common intersected points of X_1 and X_2. These points can be taken from any direction like top, bottom, left or right along with union operation.
6. **grow-diag:** In addition to neighboring points of union operation, in grow function, the aligned points are added from diagonal directions.
7. **grow-diag-final:** Considering the points in 'grow-diag', non-neighboring positioned points among words that are not aligned at least once are also included in this method.
8. **grow-diag-final-and:** Considering the points in 'grow-diag-final', non-neighboring all positioned points that are not aligned are considered in this method.

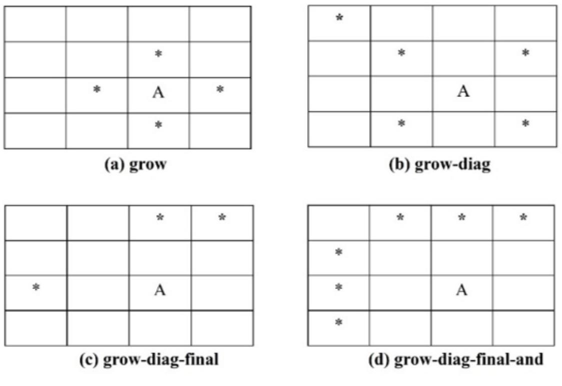
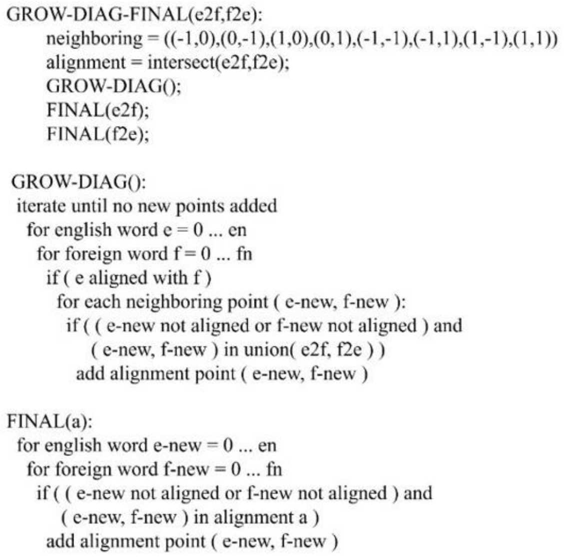


Figure 6. Different heuristic conversion functions

By the corresponding positions of alignment point 'A' which is represented by '*' in the form of a matrix is briefed in Figure 6. These eight conversions are used to design a translation model setup by using parallel corpus word alignment toolkit named as 'GIZA++' [18]. The divergent of alignment of grow symmetric function includes corresponding points from both intersection points which are unaligned and union operations. The heuristic approach of symmetric functions is elaborated in the form of pseudo code given by MOSES MT [19] and is shown in Figure 7. This pseudo code defines English 'e' as target language and 'f' denotes any foreign source language like 'Telugu'.



```
GROW-DIAG-FINAL(e2f,f2e):
    neighboring = ((-1,0),(0,-1),(1,0),(0,1),(-1,-1),(-1,1),(1,-1),(1,1))
    alignment = intersect(e2f,f2e);
    GROW-DIAG();
    FINAL(e2f);
    FINAL(f2e);

GROW-DIAG():
  iterate until no new points added
    for english word e = 0 ... en
      for foreign word f = 0 ... fn
        if ( e aligned with f )
          for each neighboring point ( e-new, f-new ):
            if ( ( e-new not aligned or f-new not aligned ) and
                ( e-new, f-new ) in union( e2f, f2e ) )
              add alignment point ( e-new, f-new )

FINAL(a):
  for english word e-new = 0 ... en
    for foreign word f-new = 0 ... fn
      if ( ( e-new not aligned or f-new not aligned ) and
          ( e-new, f-new ) in alignment a )
        add alignment point ( e-new, f-new )
```

Figure 7. Heuristic approach - Pseudo code [20]

Basically, with the use of different heuristics, the words and phrases in English to Telugu language training corpus data set are aligned. Here the compatible phrase pairs are extricated and possibilities are then allocated to the pairs. The phase pair $(e^{\wedge}, t^{\wedge})$ is treated as extricated and these are provided by words i.e., $e_1, e_2, \ldots. e_n$ in $e^{\wedge}$ are positioned points by words having in $t_1, t_2, \ldots. t_n$ in $t^{\wedge}$. This can be represented mathematically in Eq. (4).

$$\frac{count(t^{\wedge},\ e^{\wedge})}{\sum_{i=1}^{e} count(t^{\wedge}, e_i^{\wedge})} \tag{4}$$

Here $count(t^{\wedge},\ e^{\wedge})$ represent the frequency under which the phrases $e^{\wedge}$ and $t^{\wedge}$ are positioned towards each other in the corpus data set.

3.4 Reordering mechanisms

In general, there are two reordering mechanisms such as distance based and lexical based. The distance-based approach maps the words between languages according to the language structure based on distance values. The lexical-based approach positions the words according to the various directions.

3.4.1 Distance based reordering

Phrase reordering at the time of translation process is a dependent approach in terms of language conversions. The morphological structure of English is given by a sequence of subject→verb→object whereas in Telugu the sequence is subject→object→verb. Therefore, phrase reordering plays a crucial role in maintaining the standards of translation from English to Telugu sentence. This is achieved by calculating the distance mechanism while reordering of phrases and finding the estimates relying on the count of ignored words. The simple reordering mechanism is demonstrated in Figure 8.

3.4.2 Lexical based reordering

It can be done in three types of orientation methods namely monotone, discontinuous and swap. The alignment points that are positioned in the top left directions are called monotone (m). The alignment points that are positioned in the top right directions are called swap (s). The aligned points that appear neither in top right nor top left referred as discontinuous (d). The orientation methods are represented in the alignment matrix as shown in Figure 9.

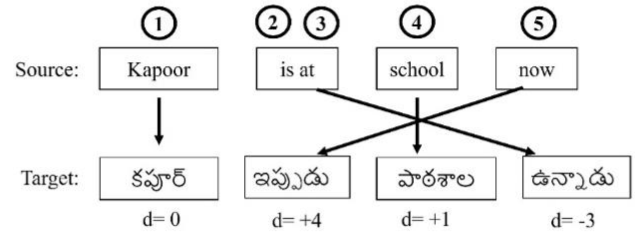


Figure 8. Reordering based on phrase distance

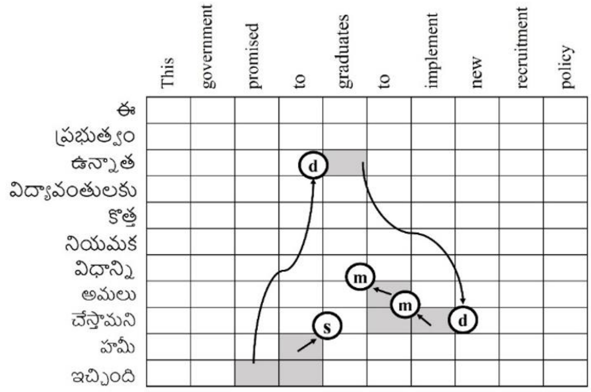


Figure 9. Reordering based on orientation

These orientation methods are represented in the form of set (e, t) having English as source and Telugu as target languages. For better understanding these orientations are represented in the form of an Eq. (5).

$$p_{orientation}((m \cup d \cup s)|e, t)) = \frac{count((m \cup d \cup s), t, e)}{\sum_{(m \cup d \cup s)} count(((m \cup d \cup s), t, e))} \tag{5}$$

where, $count(m \cup d \cup s)$ provides the frequency of monotone, discontinuous and swap orientations.

3.5 Development of reordering model

The phenomenon of reordering model can be developed in three stages, namely word, phrase and hierarchical. The word based reordering model involves in translating the context language by considering every word. While the phrase based model considers the entire phrase for translating into target language. This development of reordering is done to improve the translation quality by minimizing the error rate.

The requirement to setup a reordering model for words translation distance and lexical based reordering methods are used. From lexical method, we use MOSES lexical reordering models for exploring various combinations. This model configuration contains five parts in various aspects which are as follows [20]:

1. Modeltype - Defines the kind of model

phrase: model based on phrase

wbe: model based on word

hier: model based on hierarchical order

2. Orientation - Defines corresponding position of model

msd: monotone, swap, discontinuous

mslr: monotone, swap, discontinuous-left, discontinuous-right

leftright: left or right

monotonicity: either non-monotone or monotone

3. Directionality - Defines direction of corresponding position of model

forward: determines next phrase

backward: determines previous phrase

bidirectional: specifies both backward and forward

4. Language - Specifies the language base of model

fe: relies on both input and output languages

5. Collapsing - Specifies how to handle scores

collapseff: Scores are aligned in one direction and are combined into a single feature function.

allff: the scores can be different for each individual function

A total of 25 reordering combinations are used during the model development which contains 1 distance based and 24 lexical based aspects. MOSES toolkit which is available freely is used for training and translation process. For designing our PBRSM model, initially the default weights of the MOSES toolkit are assigned to every factor. Later, along with default parameters all other factors are analyzed in this work.

These 25 reordering factors are combined with 8 heuristic factors for translation of English to Telugu language along with n-gram models are analyzed by using TER and BLEU metrics. To maintain the standards of the language translation we used same dataset for all kinds of models namely 'wbe', 'hier' and 'phrase'.

3.6 Finding probabilities of N-gram

One of the commonly used language models is n-gram, where 'n' represents the number of words in given sentence. This method is used in predicting the next word based on the previous of it in the sentence. For example, bigram means 2 words, trigram means 3 words, 4-gram means 4 words in the given sentence. The main disadvantage of using n-gram method is it require huge size of corpus data. For every next word prediction this method verifies the entire corpus data which is complex. To handle this high volume corpus, we use Natural Language Processing (NLP) to implement n-gram approach [21, 22]. The probability of finding next word with respective to previous words in a sentence is given by Eq. (6)

$$p(w_n|w_1, w_2, w_3, \dots\dots. w_{n-1}) = \frac{count(w_1, w_2, w_3, \dots\dots. w_{n-1}, w_n)}{count(w_1, w_2, w_3, \dots\dots. w_{n-1})} \tag{6}$$

where, $w_1, w_2, w_3, \dots\dots. w_{n-1}$ is the sequence of previous words and w_n defines next word.

3.7 Dataset

Before considering the dataset, it is mandatory to preprocess the raw data. Preprocessing of the dataset can be done in the phase of cleaning. This helps in removing less useful parts present in the sentence by eliminating the commonly use words, duplicates, empty spaces, special characters and other unwanted data.

Table 2. Summary of dataset [23]

Type	**Sentence Count**		**Number of Tokens**			
	English	**Telugu**	**3-gram**	**4-gram**	**5-gram**	**6-gram**
Total sentences	2179	2179	651	1304	2180	3270
Training data	1743	1743	520	1043	1744	2616
Testing data	436	436	131	261	436	654

Then, we combined two corpus data sets in analyzing, training and validation for English to Telugu translation in this research. Initial set of data gathered from "indian-parallel-corpora" dataset [23]. The gathered data is updated by eliminating noises, misplaced usual phrases and disparities in grammatical syntaxes. By performing few initial procedures like assigning tokes, true casing with the help of default tools provided by the MOSES toolkit. The aim of this work is to perform the effective use of the PBRSM modules without expanding the hardware components. The probabilities in development, analyzing and training corpus data sets for all initial model is shown in Table 2. In every data set the count of phrase is the count of sentences and different words is represented as its vocabulary. This dataset is divided into two phases for training (80%) and testing (20%) purpose. It contains a greater number of tokens for 6-grams and a smaller number of tokens for 3-gram.

4. RESULTS AND DISCUSSION

4.1 Evaluation metrics

The level of quality translation relies on the accuracy and Excellency of the target language. This is acquired from finding resemblance among human produced source conversion and the system produced target language.

4.1.1 Bilingual Evaluation Understudy (BLEU)

When the n-gram precision score is studied, n-gram similar matches can be calculated [24] The BLEU value of entire data is given through mean of BLEU value of independent group of phrases. This BLEU value can be obtained through the Eq. (7). The scores obtained are considered as most accurate while the comparisons are done at corpus instead of paragraph level.

$$BLEU = brevity\ penalty * \exp \sum_{i=1}^{n} \lambda_i \log precision \tag{7}$$

where, the '$penalty\ of\ brevity$' represents the exact count of words in a given input sentence. Weights 'λ_i' is used for different precisions are initially set to 1. The better translation is considered when higher the BLEU score that the system produced in reference to translation on showing high correlation.

4.1.2 Translation Edit Rate (TER)

This metric relies on distance based metric [25]. This is used in finding the least count of substitutes which are required in the MT produced conversions for making it easier to the target conversion. The complex operations like shift, delete, insert and substitute that are utilized in editing purpose. The score TER is given by the Eq. (8).

$$TER = \frac{Number\ of\ edits\ of\ (I\ U\ D\ \ U\ S_h\ U\ S_u)}{Average\ number\ of\ word\ references} \tag{8}$$

Here, I denotes insert operation, D denotes deletion operation, S_h refers shift operation and S_u denotes substitute operation. Higher the TER value represents a greater number of editing operations. These are needed in MT system conversion result correctly and same as that of the referred conversion. Thus, the MT system conversion having less TER value is referred as good conversion.

4.2 Performance of reordering models

The main aim of Table 3 to Table 5, is to identify the accurate settings of the alignment of words in heuristic, language and reordering method between English to Telugu translation. Nearly 4,358 sentences from English and Telugu languages are considered for measuring the performance of PBRSM model. In such case the quality of translation can also be evaluated. Using different combinations of language, reordering and word alignment heuristic models, the metrics of BLEU and TER scores of PBRSM are evaluated. The results generated through our proposed model including grow-diag, grow-diag-final-and, insertion and union heuristic word alignments are shown in the format of tables are represented. Initially, we have received BLEU and TER scores for 3-gram, 4-gram, 5-gram and 6-gram. Out of which, 6-gram has attained higher scores. Therefore, experimentally we have shown the outcomes of 6-gram instead of lower size grams.

The TER and BLEU metric scores of lexical and distance based models among 'intersection', 'union', 'grow' and 'grow-diag-final-and' orientations on 6-gram of PBRSM model are mentioned in Table 3. For intersection, the model 'phrase→mslr→bidirectional→fe→allff' has received best TER and BLEU scores with 62.01 and 29.87. For union, the model 'phrase→mslr→bidirectional→fe→allff' has attained least TER score of 58.59 and the model 'phrase→mslr→bidirectional→fe→collapseff' has received best BLEU score with 30.76. For grow, the model 'phrase→mslr→bidirectional→fe→allff' has attained least TER score with 57.43 and the model 'phrase→mslr→bidirectional→fe→collapseff' has received best BLEU score with 31.69. For grow-diag-final-and, the model 'phrase→mslr→bidirectional→fe→allff' has received least TER score with 55.6 and for BLEU the model 'phrase→mslr→bidirectional→fe→collapseff' has attained best score with 32.94.

The outcomes of WBRSM on various heuristics such as intersection, union, grow, grow-diag-final-and are shown in Table 4. For intersection, the model 'wbe→monotonicity→ bidirectional→fe→allff' showed least TER score with 68.26 and the model 'wbe→monotonicity→ bidirectional→fe→collapseff' achieved best BLEU score with 25.97. For union, grow and grow-diag-final-and, the model 'wbe→mslr→bidirectional→fe→allff' has received least TER scores of 66.92, 66.08, 64.91 and in case of BLEU the model 'wbe→mslr→bidirectional→fe→collapseff' achieved best score of 25.79, 26.56, 26.97 respectively.

From Table 5, the comparisons are carried out on the model HBRSM. For intersection, union, grow and grow-diag-final-and, the model 'hier→mslr→bidirectional→fe→allff' has received least TER score of 72.52, 70.91, 70.39, 69.89 and the model 'hier→mslr→bidirectional→fe→collapseff' has attained best BLEU score of 20.42, 20.63, 20.74, 20.8 respectively.

Table 3. TER and BLEU scores for 6-gram of PBRSM on various heuristics

Serial No	Type	Reordering Model	Intersection		Union		Grow		Grow-diag-final-and	
			TER	BLEU	TER	BLEU	TER	BLEU	TER	BLEU
1	Phrase based	phrase→msd→forward→fe→allff	65.31	27.04	62.09	27.93	65.31	27.04	62.09	27.93
2		phrase→msd→forward→fe→collapseff	65.31	27.04	62.09	27.93	65.31	27.04	62.09	27.93
3		phrase→msd→backward→fe→allff	63.95	24.83	60.73	25.72	63.95	24.83	60.73	25.72
4		phrase→msd→backward→fe→ collapseff	63.95	24.83	60.73	25.72	63.95	24.83	60.73	25.72
5		phrase→msd→bidirectional→fe→allff	62.26	21.33	59.04	22.22	62.26	21.33	59.04	22.22
6		phrase→msd→bidirectional→fe→ collapseff	63.84	18.24	60.62	19.13	63.84	18.24	60.62	19.13
7		phrase→mslr→forward→fe→allff	68.39	21.59	65.17	22.48	68.39	21.59	65.17	22.48
8		phrase→mslr→forward→fe→ collapseff	68.39	21.59	65.17	22.48	68.39	21.59	65.17	22.48
9		phrase→mslr→backward→fe→allff	65.34	26.37	62.12	27.26	65.34	26.37	62.12	27.26
10		phrase→mslr→backward→fe→collapseff	65.34	26.37	62.12	27.26	65.34	26.37	62.12	27.26
11		phrase→mslr→bidirectional→fe→allff	62.01	29.36	58.79	30.25	62.01	29.36	58.79	30.25
12		phrase→mslr→bidirectional→fe→ collapseff	63.31	29.87	60.09	30.76	63.31	29.87	60.09	30.76
13		phrase→leftright→forward→fe→allff	64.43	23.98	61.21	24.87	64.43	23.98	61.21	24.87
14		phrase→leftright→forward→fe→colapseff	64.43	23.98	61.21	24.87	64.43	23.98	61.21	24.87
15		phrase→leftright→backward→fe→allff	63.46	24.73	60.24	25.62	63.46	24.73	60.24	25.62
16		phrase→leftright→backward→fe→colapseff	63.46	24.73	60.24	25.62	63.46	24.73	60.24	25.62
17		phrase→leftright→bidirectional→fe→allff	64.29	24.38	61.07	25.27	64.29	24.38	61.07	25.27
18		phrase→leftright→bidirectional→fe→colapseff	63.71	24.56	60.49	25.45	63.71	24.56	60.49	25.45
19		phrase→monotonicity→forward→fe→allff	63.68	25.11	60.46	26	63.68	25.11	60.46	26
20		phrase→monotonicity→forward→fe→ colapseff	63.68	25.11	60.46	26	63.68	25.11	60.46	26
21		phrase→monotonicity→backward→fe→ allff	63.5	24.88	60.28	25.77	63.5	24.88	60.28	25.77
22		phrase→monotonicity→backward→fe→ colapseff	63.5	24.88	60.28	25.77	63.5	24.88	60.28	25.77
23		phrase→monotonicity→bidirectional→fe→ allff	63.82	25	60.6	25.89	63.82	25	60.6	25.89
24		phrase→monotonicity→bidirectional→fe→ colapseff	64.58	24.04	61.36	24.93	64.58	24.04	61.36	24.93
25	Distance based	Distance	62.92	25.28	59.7	26.17	62.92	25.28	59.7	26.17

Table 4. TER and BLEU scores for 6-gram of WBRSM on various heuristics

Serial No	Type	Reordering Model	Intersection		Union		Grow		Grow-diag-final-and	
			TER	BLEU	TER	BLEU	TER	BLEU	TER	BLEU
1	Phrase based	phrase→msd→forward→fe→allff	72.69	22.47	70.22	22.96	72.69	22.47	70.22	22.96
2		phrase→msd→forward→fe→collapseff	72.69	22.47	70.22	22.96	72.69	22.47	70.22	22.96
3		phrase→msd→backward→fe→allff	71.33	20.26	68.86	20.75	71.33	20.26	68.86	20.75
4		phrase→msd→backward→fe→ collapseff	71.33	20.26	68.86	20.75	71.33	20.26	68.86	20.75
5		phrase→msd→bidirectional→fe→allff	69.64	16.76	67.17	17.25	69.64	16.76	67.17	17.25
6		phrase→msd→bidirectional→fe→ collapseff	71.22	13.67	68.75	14.16	71.22	13.67	68.75	14.16
7		phrase→mslr→forward→fe→allff	75.77	17.02	73.3	17.51	75.77	17.02	73.3	17.51
8		phrase→mslr→forward→fe→ collapseff	75.77	17.02	73.3	17.51	75.77	17.02	73.3	17.51
9		phrase→mslr→backward→fe→allff	72.72	21.8	70.25	22.29	72.72	21.8	70.25	22.29
10		phrase→mslr→backward→fe→collapseff	72.72	21.8	70.25	22.29	72.72	21.8	70.25	22.29
11		phrase→mslr→bidirectional→fe→allff	69.39	24.79	66.92	25.28	69.39	24.79	66.92	25.28
12		phrase→mslr→bidirectional→fe→ collapseff	70.69	25.3	68.22	25.79	70.69	25.3	68.22	25.79
13		phrase→leftright→forward→fe→allff	71.81	19.41	69.34	19.9	71.81	19.41	69.34	19.9
14		phrase→leftright→forward→fe→colapseff	71.81	19.41	69.34	19.9	71.81	19.41	69.34	19.9
15		phrase→leftright→backward→fe→allff	70.84	20.16	68.37	20.65	70.84	20.16	68.37	20.65
16		phrase→leftright→backward→fe→colapseff	70.84	20.16	68.37	20.65	70.84	20.16	68.37	20.65
17		phrase→leftright→bidirectional→fe→allff	71.67	19.81	69.2	20.3	71.67	19.81	69.2	20.3
18		phrase→leftright→bidirectional→fe→colapseff	71.09	19.99	68.62	20.48	71.09	19.99	68.62	20.48
19		phrase→monotonicity→forward→fe→allff	71.06	20.54	68.59	21.03	71.06	20.54	68.59	21.03
20		phrase→monotonicity→forward→fe→ colapseff	71.06	20.54	68.59	21.03	71.06	20.54	68.59	21.03
21		phrase→monotonicity→backward→fe→ allff	70.88	20.31	68.41	20.8	70.88	20.31	68.41	20.8
22		phrase→monotonicity→backward→fe→ colapseff	70.88	20.31	68.41	20.8	70.88	20.31	68.41	20.8
23		phrase→monotonicity→bidirectional→fe→ allff	68.26	25.83	68.73	20.92	68.26	25.83	68.73	20.92
24		phrase→monotonicity→bidirectional→fe→ colapseff	68.96	25.97	69.49	19.96	68.96	25.97	69.49	19.96
25	Distance based	Distance	70.3	20.71	67.83	21.2	70.3	20.71	67.83	21.2

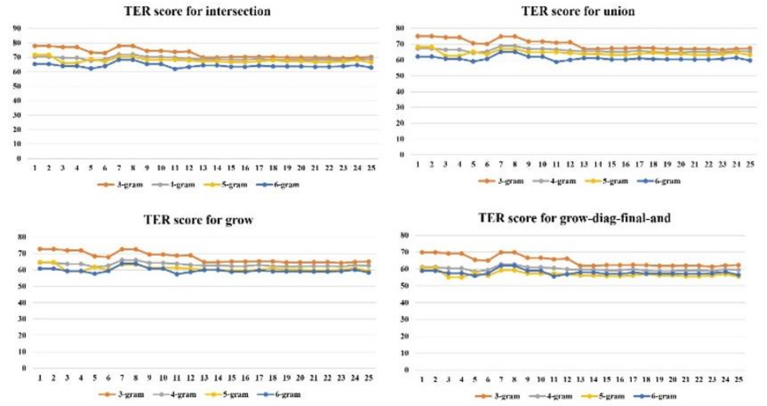


Figure 10. TER score for PBRSM with various orientations

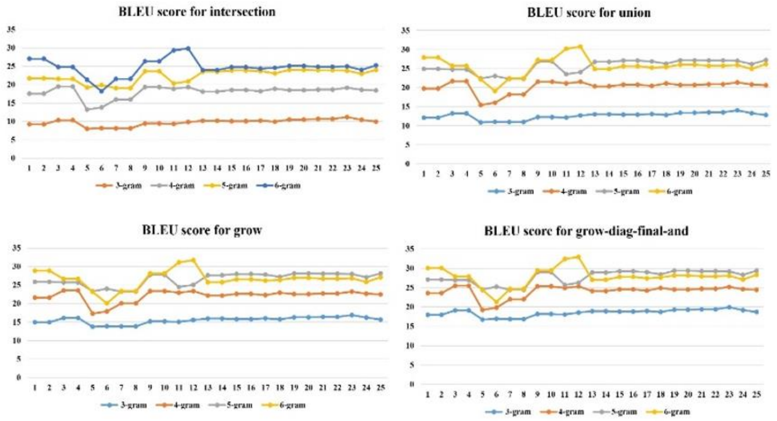


Figure 11. BLEU score for PBRSM with various orientations

Table 5. TER and BLEU scores for 6-gram of HBRSM on various heuristics

Serial No	Type	Reordering Model	Intersection		Union		Grow		Grow-diag-final-and	
			TER	BLEU	TER	BLEU	TER	BLEU	TER	BLEU
1	Phrase based	phrase→msd→forward→fe→allff	75.82	17.59	74.21	17.8	73.69	17.91	73.19	17.97
2		phrase→msd→forward→fe→collapseff	75.82	17.59	74.21	17.8	73.69	17.91	73.19	17.97
3		phrase→msd→backward→fe→allff	74.46	15.38	72.85	15.59	72.33	15.7	71.83	15.76
4		phrase→msd→backward→fe→ collapseff	74.46	15.38	72.85	15.59	72.33	15.7	71.83	15.76
5		phrase→msd→bidirectional→fe→allff	72.77	11.88	71.16	12.09	70.64	12.2	70.14	12.26
6		phrase→msd→bidirectional→fe→ collapseff	74.35	8.79	72.74	9	72.22	9.11	71.72	9.17
7		phrase→mslr→forward→fe→allff	78.9	12.14	77.29	12.35	76.77	12.46	76.27	12.52
8		phrase→mslr→forward→fe→ collapseff	78.9	12.14	77.29	12.35	76.77	12.46	76.27	12.52
9		phrase→mslr→backward→fe→allff	75.85	16.92	74.24	17.13	73.72	17.24	73.22	17.3
10		phrase→mslr→backward→fe→collapseff	75.85	16.92	74.24	17.13	73.72	17.24	73.22	17.3
11		phrase→mslr→bidirectional→fe→allff	72.52	19.91	70.91	20.12	70.39	20.23	69.89	20.29
12		phrase→mslr→bidirectional→fe→ collapseff	73.82	20.42	72.21	20.63	71.69	20.74	71.19	20.8
13		phrase→leftright→forward→fe→allff	74.94	14.53	73.33	14.74	72.81	14.85	72.31	14.91
14		phrase→leftright→forward→fe→colapseff	74.94	14.53	73.33	14.74	72.81	14.85	72.31	14.91
15		phrase→leftright→backward→fe→allff	73.97	15.28	72.36	15.49	71.84	15.6	71.34	15.66
16		phrase→leftright→backward→fe→colapseff	73.97	15.28	72.36	15.49	71.84	15.6	71.34	15.66
17		phrase→leftright→bidirectional→fe→allff	74.8	14.93	73.19	15.14	72.67	15.25	72.17	15.31
18		phrase→leftright→bidirectional→fe→colapseff	74.22	15.11	72.61	15.32	72.09	15.43	71.59	15.49
19		phrase→monotonicity→forward→fe→allff	74.19	15.66	72.58	15.87	72.06	15.98	71.56	16.04
20		phrase→monotonicity→forward→fe→ colapseff	74.19	15.66	72.58	15.87	72.06	15.98	71.56	16.04
21		phrase→monotonicity→backward→fe→ allff	74.01	15.43	72.4	15.64	71.88	15.75	71.38	15.81
22		phrase→monotonicity→backward→fe→ colapseff	74.01	15.43	72.4	15.64	71.88	15.75	71.38	15.81
23		phrase→monotonicity→bidirectional→fe→ allff	74.33	15.55	72.72	15.76	72.2	15.87	71.7	15.93
24		phrase→monotonicity→bidirectional→fe→ colapseff	75.09	14.59	73.48	14.8	72.96	14.91	72.46	14.97
25	Distance based	Distance	73.43	15.83	71.82	16.04	71.3	16.15	70.8	16.21

From the Table 3 to Table 5, it can be concluded that we have shown the results of 6-gram model. Where in the remaining 3-gram,4-gram and 5-gram are done experimentally to avoid the ambiguity on different type of grams. Hence, the highest score is attained for 6-gram and is shown in a graphical representation. In case of BLEU metric, the PBRSM showed accurate results for finding the similarities between source and target languages from 3-gram to 6-gram. The results of PBRSM model with parameters of 'insertion', 'union', 'grow', 'grow-diag-final-and' word alignment heuristics are represented in graphical in Figure 10 for the evaluation of TER score. The same parameters are considered in case of evaluating BLEU score are shown in a graphical presentation in Figure 11. In Figure 10 and Figure 11, *x-axis* represents a serial reordering model in the resulting tables and the metrics of BLEU and TER scores are shown in y-*axis* respectively. Thus, the outcomes obtained are known evaluated individually to find the efficiency of language, reordering and word alignment models of PBRSM in terms of quality during translation. In case of testing the PBRSM model performance at different corpus data is used therefore, low BLEU and TER scores are obtained. By considering all the orientations with n-gram models on phrase, word and hierarchical, it is observed that PBRSM achieved least error rate scores of TER and

BLEU metrics. From these it can be observed that our proposed PBRSM model have shown promising results with low error rate when there is an increase in 3-to-6 grams for concerned orientations.

From Figure 10, it is observed that different orientations of PBRSM model with TER metric is analyzed. From this we can conclude that, for the orientations of 'intersection', 'union' and 'grow' received better results for 6-gram compared to other n-grams used in this work. For the 'grow-diag-final-and' orientation obtained good outcome for 5-gram when compared to that of 6-gram. From Figure 11, it is observed that the BLEU metric outcome is higher for 6-gram for 'intersection' orientation when compared with other n-grams. The 'union', 'grow' and 'grow-diag-final-and' orientations obtained higher BLEU score for 5-gram sentence than 6-gram.

5. CONCLUSION AND FUTURE SCOPE

The quality of language translation depends upon the performance of reordering models like Lexical and Distance based for n-gram models. For translating English sentence into Telugu language in this work, we used PBRSM, WBRSM and HBRSM lexical based models. According to this model performance, it is observed that PBRSM model is considered as the best model. It achieved least error rate of TER with 62.01 and higher value of BLEU with 29.07 metric value for the 6-gram model 'phrase→mslr→bidirectional→fe→allff' in all orientations. However, for the WBRSM and HBRSM performances, it is observed that in 6-gram model different orientations showed different models. Due to this, it becomes very complex to identify the best model among various orientations which leads to ambiguity. It requires huge corpus dataset thereby reducing the quality and standard of the translated sentence. Our model is best suited in private and public sectors in understands the foreign language in their regional language. Hence, in future work, our aim is to handle complex parallel corpus data using Deep Learning methods among all reordering lexical models. Through deep learning models, we can convert one language to another through image processing methods by deeply understanding the character recognition in the given corpus data.

ACKNOWLEDGMENT

We gratefully acknowledge the AI Center at Norwich University for providing the research environment and resources essential for the execution of this work. This research was made possible through the grant "Advanced Computing through Experiential Education" awarded by the Department of Education (award number: P116Z220106).

DATA AVAILABILITY

The data used to support the findings of this study is freely available at "https://github.com/joshua-decoder/indian-parallel-corpora/tree/master/te-en/tok".

CONFLICT OF INTEREST

The authors declare that there are no conflicts of interest regarding the publication of this paper.

REFERENCES

[1] Censusindia.gov.in. Census of India 2011-LANGUAGE ATLAS-INDIA-India. https://censusindia.gov.in/nada/index.php/catalog/42561.

[2] Kanwal, S. (2011). India-most common languages 2011. https://www.statista.com/statistics/616508/most-common-languages-india.

[3] Ethnologue Telugu | Ethnologue Free. https://www.ethnologue.com/language/tel.

[4] Udupa, U.R., Faruquie, T.A. (2004). An English-Hindi statistical machine translation system. In International Conference on Natural Language Processing, Hainan Island, China, pp. 254-262. https://doi.org/10.1007/978-3-540-30211-7_27

[5] Suryakanthi, T., Prasad, S.V.A.V., Prasad, T.V. (2013). Translation of pronominal anaphora from English to Telugu language. International Journal of Advanced Computer Science and Applications, 4(4): 75-79. http://dx.doi.org/10.14569/IJACSA.2013.040413

[6] Natu, I., Iyer, S., Kulkarni, A., Patil, K., Patil, P. (2018). Text translation from Hindi to English. In Advances in Computing and Data Sciences: Second International Conference, ICACDS 2018, Dehradun, India, April 20-21, 2018, Revised Selected Papers, Part I 2, Dehradun, India, pp. 481-488. https://doi.org/10.1007/978-981-13-1810-8

[7] Prasad, T.V., Muthukumaran, G.M. (2013). Telugu to English translation using direct machine translation approach. International Journal of Science and Engineering Investigations, 2(12): 25-32.

[8] Nair, L.R., David Peter, S. (2012). Machine translation systems for Indian languages. International Journal of Computer Applications, 39(1): 0975-8887.

[9] Dungarwal, P., Chatterjee, R., Mishra, A., Kunchukuttan, A., Shah, R., Bhattacharyya, P. (2014). The IIT Bombay Hindi-English translation system at WMT 2014. In Proceedings of the Ninth Workshop on Statistical Machine Translation, Baltimore, Maryland USA, pp. 90-96.

[10] Och, F.J., Ney, H. (2003). A systematic comparison of various statistical alignment models. Computational linguistics, 29(1): 19-51. https://doi.org/10.1162/089120103321337421

[11] Och, F.J., Ney, H. (2004). The alignment template approach to statistical machine translation. Computational linguistics, 30(4): 417-449. https://doi.org/10.1162/0891201042544884

[12] Dutta Chowdhury, K., Hasanuzzaman, M., Liu, Q. (2018). Multimodal neural machine translation for low-resource language pairs using synthetic data. Association for Computational Linguistics (ACL).

[13] Reddy, M.V., Hanumanthappa, M. (2013). Indic language machine translation tool: English to Kannada/Telugu. In Multimedia Processing, Communication and Computing Applications: Proceedings of the First International Conference, Springer New Delhi, pp. 35-49. https://doi.org/10.1007/978-81-322-1143-3

[14] Lingam, K., Lakshmi, E.R., Theja, L.R. (2014). Rule-based machine translation from English to Telugu with emphasis on prepositions. In 2014 First International Conference on Networks & Soft Computing (ICNSC2014), Guntur, India, pp. 183-187.

https://doi.org/10.1109/CNSC.2014.6906669

[15] Suryakanthi, T., Sharma, K. (2015). Discourse translation from English to Telugu. In Proceedings of the third international symposium on women in computing and informatics, Kochi, India, pp. 222-227. https://doi.org/10.1145/2791405.2791459

[16] Raju, B.N., Raju, M.B., Satyanarayana, K.V.V. (2021). Effective preprocessing based neural machine translation for English to Telugu cross-language information retrieval. IAES International Journal of Artificial Intelligence, 10(2): 306. https://doi.org/10.11591/ijai.v10.i2.pp306-315

[17] Babhulgaonkar, A., Sonavane, S. (2022). Empirical analysis of phrase-based statistical machine translation system for English to Hindi language. Vietnam Journal of Computer Science, 9(2): 135-162. https://doi.org/10.1142/S219688882250004X

[18] Moses-Moses/Background. (2023). http://www2.statmt.org/moses/?n=Moses.Backgrouund.

[19] Moses-FactoredTraining/RunGIZA. (2023). http://www2.statmt.org/moses/?n=FactoredTraining.RunGIZA.

[20] Moses-FactoredTraining/AlignWords. (2023). http://www2.statmt.org/moses/?n=FactoredTraining.AlignWords.

[21] Moses-FactoredTraining/BuildReorderingModel. (2023). http://www2.statmt.org/moses/?n=FactoredTraining.BuildReorderingModel.

[22] Srinidhi, S. (2023). Understanding Word N-grams and N-gram Probability in Natural Language Processing. https://towardsdatascience.com/understanding-word-n-grams-and-n-gram-probability-in-natural-language-processing-9d9eef0fa058.

[23] Joshua. Indian-parallel-corpora/te-en/tok. (2023). https://github.com/joshua-decoder/indian-parallel-corpora/tree/master/te-en/tok.

[24] Papineni, K., Roukos, S., Ward, T., Zhu, W.J. (2002). Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 311-318. https://aclanthology.org/P02-1040.pdf.

[25] Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers, Cambridge, Massachusetts, USA, pp. 223-231. https://aclanthology.org/2006.amta-papers.25.