International Journal of Future Generation Communication and Networking
Vol. 13, No. 3, (2020), pp. 1324–1337

Optimization of Phrase Table in Statistical Machine Translation Using Linguistic Features

*Arun Babhulgaonkar[1], Shefali Sonavane[2]

[1]*Department of IT, Dr. Babasaheb Ambedkar Technological University, Lonere, India.*
arbabhulgaonkar@dbatu.ac.in
[2]*Department of IT, Walchand College of Engineering, Sangli, India.*
shefali.sonavane@walchandsangli.ac.in

Abstract

Phrase based machine translation is a statistical approach of machine translation that heavily depends on the quality and size of the phrase table. During the training process, phrase table accumulates many useful as well as useless phrase pairs obtained from word aligned parallel corpus. The huge size of phrase table due to useless, spurious and redundant phrase pairs demands more memory and unnecessary processor time. To accommodate the translation model on handheld devices where memory is a scarce resource, some research has been done to minimize the size of the phrase table. However, most of the existing approaches remove the useless phrase pairs using hard rule based translation probability values. The rigid constraints in these methods use only a single probability value factor and do not consider composite factors. This paper proposes a machine learning based classification framework to filter the useless phrases from the phrase table using multiple linguistic and statistical features. Results obtained show that, the proposed framework efficiently removes 67% of the phrase pairs while keeping acceptable quality of translation.

Keywords: *PBSMT, Classification, Translation Table Pruning, Syntactic Constraints, Word Alignment, BLEU, TER*

1. Introduction

Phrase based statistical machine translation (PBSMT) system can be utilized in low resource conditions to produce neural machine translation comparable translation results. PBSMT uses any arbitrary sequence of words called as phrase as a translation unit [6]. A phrase need not to be a linguistic entity. Figure 1 illustrates the phrase based model. During training of the model, the input source and target parallel sentences are partitioned into a sequence of phrases and phrase alignments are obtained. Obtained phrase pairs are then stored in a phrase table along with their probability values. During translation, using phrase table, for the given input source sentence target sentence is constructed through a generative process.

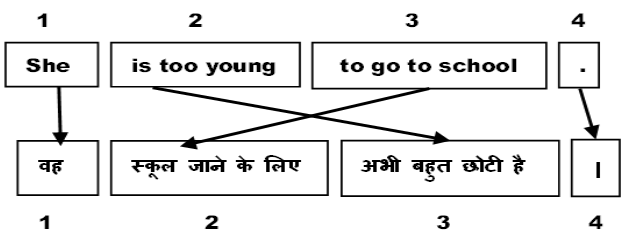

Figure 1. Segmentation of a Sentence into Phrases

ISSN: 2233-7857 IJFGCN

1.1. Learning a Phrase Table

The phrase table works as the central knowledge source for the PBSMT system. Usually, phrase table is obtained from the word alignments. In Figure 2, the alignments between the words of an English sentence *"Sachin was honored with the Bharatratna award"* and *Hindi* sentence *"**सचिन को भारतरत्न पुरस्कार से सम्मानित किया गया**"* are represented by dark cells in the alignment matrix. Consistent phrases may be longer or shorter in length. The problem of small phrase size is that it will create many phrases that leads to a huge phrase table. However, it is observed that the most of the long length phrases in the training data never appear in test data again. Another cause of huge size of phrase table is unaligned words like *"the"* on the source side of the translation process as given in Figure 2.

Sachin	सचिन
Sachin	सचिन को
was honored with	से सम्मानित किया
was honored with	से सम्मानित किया गया
honored with	से सम्मानित
with the Bharatratna award	भारतरत्न पुरस्कार से
Bharatratna award	भारतरत्न पुरस्कार
the Bharatratna award	भारतरत्न पुरस्कार
was honored with the Bharatratna award	भारतरत्न पुरस्कार से सम्मानित किया गया

Figure 2. Phrase Pairs Extracted from the Word Alignment

1.2. Problem of Size of the Phrase Table

8 *Hindi* words and 7 English words are matched by 5 alignment points in the given example in Figure 2. From such a small sentence pair 9 phrase pairs are extracted. For any arbitrary phrase segmentation, all phrase pairs consistent with the word alignments need to be considered. Generally, the number of phrases extracted is roughly quadratic in the number of words. The size of phrase table requires Gigabyte to Terabyte of memory for large parallel corpora with millions of sentences. This may become too much to accommodate in the working memory of computer system. Due to this, estimation of phrase translation probabilities and its use to translate new sentences becomes difficult. For portable devices like mobiles, PDAs, etc., CPU performance and memory shortage are the biggest limitations in producing a real time and acceptable quality translation output. The big size of phrase table demands more storage which increases the memory cost of these devices. This limitation of PBSMT system motivates to focus on the translation model pruning as a research problem of this paper. In reality, many parallel phrases are redundant and do not contribute in actual language translation process. Identification and removal of such phrases from the phrase table leads to reduction in phrase table size and also improves the processing speed. It is observed that, typically two types of phrase pairs cause the unnecessary increase in the size of the phrase table. First

ISSN: 2233-7857 IJFGCN

type is phrase pairs generated in 1-to-many scenario in which a distinct source phrase is associated with many translation options. Many of these options can be safely discarded. The second type is phrase pairs generated in 1-to-few scenario in which some distinct peculiar source phrases in reality possesses only one or two translation options. In such scenario, many phrases are extracted by mistake because of wrong alignments. Threshold based pruning methods put some cut off value to remove overloaded translation options, but these methods always fail to remove the phrases generated in second scenario. Fisher's significance test, "Noise', relative entropy is also used for pruning the weakly associated phrase pairs [15]. However, this kind of approaches may unnecessarily remove many useful phrase pairs such as named entities, date and time related quantities which occur very rarely in the corpus. The drawback of all existing methods is that they use statistical or syntactic information in isolation. This makes phrase table pruning still an open research problem in the field of SMT. The structural difference between two languages of different origin such as English and *Hindi* can provide additional useful information for identification of useless phrases to be removed. This paper tries to propose a classification based framework to identify and retain only useful phrase pairs. For classification, many heuristic measures used in literature for pruning purpose are encoded as features instead of using them as hard rules. Many features such as statistical value, syntactic information and Marker word information are used together to identify useless phrases and remove them from the phrase table.

1.3. Proposed Machine Learning Based Pruning Framework

As the requirement of the classification approach, all the available phrases in the phrase table are divided into two classes based on their usefulness criteria as potentially more useful and less useful. Figure 3 shows the translation model pruning framework. It illustrates how the classifier is integrated in the process for getting a pruned translation model. Once the basic translation model is obtained from word alignments, all of the parameters of the system are tuned with a development dataset. Minimum Error Rate Training (MERT) technique is used for tuning the parameters of the system. This tuned translation model is then used to trace the decoding path of the sentences in the development dataset to get classifier training data. Finally, pruned translation model is constructed after applying the classifier to remove potentially less useful phrases in the phrase table.

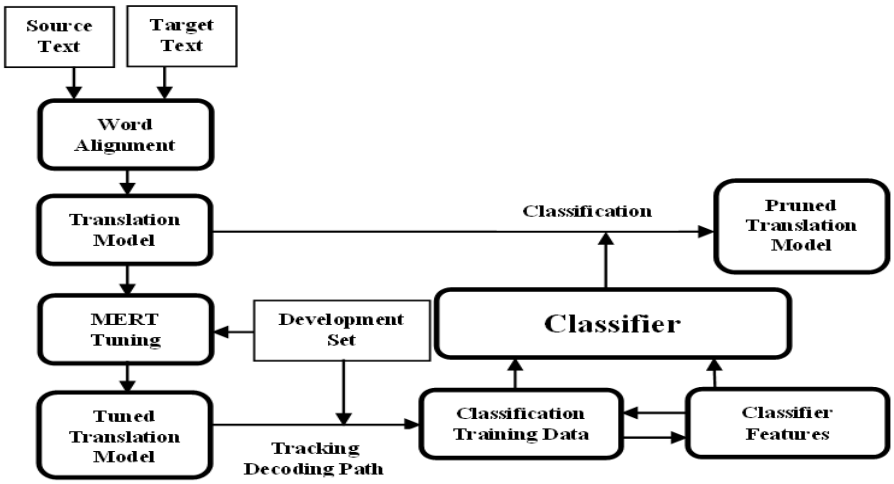

Figure 3. Block Diagram of Proposed Pruning Framework

2. Related Work

ISSN: 2233-7857 IJFGCN

Most simple method of pruning the unnecessary phrases from phrase translation table is based on usage statistics [3]. In this method, the developed translation system is applied on a large amount of development dataset and relative importance of each phrase pair is calculated by obtaining its actual usage in the process of decoding. A similar approach is proposed in paper [14] which suggested to extract only those phrase pairs which produce high scoring metric for translations. Johnson et al. [5] used simpler independence measure like 'p-value' associated with phrase pair co-occurrence for pruning purpose. They used a filtering algorithm to retain only those phrase pairs which will pass a significance test. The significance test finds the possibility of co-occurrences of a phrase pair again in a text is low or high. Another statistical independence measure called as Noise is used by authors in paper [11]. They utilized Noise filtering criterion to divide the bi-phrase tables in several sub-tables according to their phrase size complexity. Authors argued that use of Noise as a filtering criterion is more effective than use of p-value as per as translation quality is concerned. Some researchers also proposed to use some machine learning techniques for table pruning purpose. Ling et al. [8] utilized relative entropy for removing the long length phrase pairs that can be constructed by small length phrase pairs with equal probability and translation quality. Wang et al. [13] combined the relative entropy and significance approach for pruning. All the approaches mentioned earlier ignore the important attribute like language syntactic features of phrases. Due to this sometimes high chances are there that these approaches discard the potentially useful phrases because of low probability associated with them. Authors in paper [6] used strict syntactic constraint for pruning. Due to this stringent criteria, many phrase pairs get removed from table but some of these phrase pairs may be well translated pairs. Cao et al. [1] used relaxed syntactic constraints in the source language phrase and proved that this is not harming the translation quality. However, the target language syntactic information for pruning is not explored and still there is a scope to use it. Kavitha et al. [7] introduced the use of classification for the term pairs validation problem for building translation lexicon. They utilized co-occurring frequencies of source and target terms as a feature for classification. The use of classifier framework for phrase table pruning proposed in this paper is inspired from the idea presented in papers [9][12]. The authors in [12] used single-class SVM to identify good phrase pairs. They used the Mapping Convergence algorithm and single-class SVM to identify useful and useless phrase pairs. The idea of using single-class classification technique for distinguishing positive and negative instances is very complicated. The extensions in the proposed framework are: 1) Pruning the translation model while keeping comparable translation quality. 2) Use of various linguistic and statistical values as classifier features 3) Generation of classifier training data automatically from the development dataset.

3. Classifier Construction

3.1. Training Set for the Classifier

How positive and negative class examples required for training of the classifier are obtained is described next. First, using source sentences in the development dataset, development phrase table, D={Dev-TT}, is obtained from the original phrase table, O={TT}. Then, the search graph is traced during decoding of the sentences in the development dataset using tuned translation model and phrase pairs present in the decoding lattice are recorded. All the phrase pairs in Dev-TT are divided into three categories: 'finally-used', 'considered but not finally used' and 'not considered'. This division of phrase pairs is inspired from the usage statistics concept as explained in [3]. Out of these three categories, the phrase pairs marked as 'finally-used' are actually considered and included in the final translation, hence used as positive class examples and the phrase pairs marked as 'not-considered' are used as negative class examples. Remaining phrase pair sets can be defined as follows:

ISSN: 2233-7857 IJFGCN

B={TT-In-Beam}: the translation table in beam containing only those phrase pairs which have been considered at least once in the construction of translation lattice during decoding.

U={Used-TT}: table contains phrase pairs which are actually utilized in the final translations. U is subset of B as U contains only those phrase pairs in B that are used in the best translations.

N={TT-Out-Of-Beam}: table contains those phrase pairs that are discarded by the beam search. (N=D-B).

Elements in set U are considered as positive samples and elements in set N are considered as negative samples. Figure 4 shows all the constructed sets of phrase pairs using Venn diagram. The closed interval B-U contains the phrase pairs which are scrutinized during decoding process but are not included in the final translation.

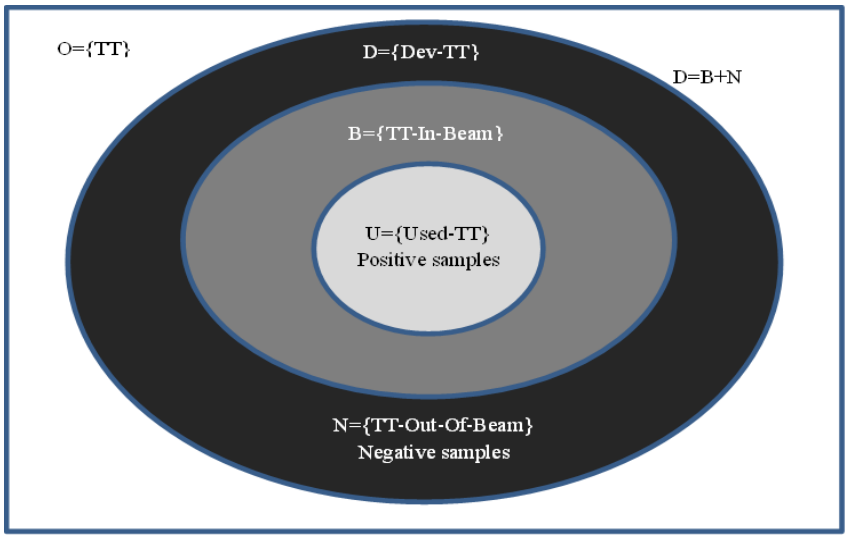

Figure 4. Positive and Negative Class Training Dataset

3.2. Features for the Classifier

Idea of how to extract the classifier features is inspired by previous studies. Overall four types of features are considered for building the classifier: syntactic constraints, bidirectional translation probabilities, significance value and length ratio. From the phrase table, bidirectional phrase translation probabilities and bidirectional lexicalized phrase translation probabilities are obtained. From the reordering table, the reordering probabilities are obtained. The syntactic constraints are obtained from syntactic parse tree. If the phrase span is part of a single sub-tree of parse tree, it is considered as a syntactic phrase and is used as a syntactic constraint in paper [1]. If the source phrase is not a syntactic phrase and the first or last word in it is not aligned, then authors discarded such phrase pair. For example, in the phrase pair "does not: नहीं" in the sentence pair "Arun does not live here : अरुण यहां नहीं रहते", first word "does" is having no equivalent in *Hindi* language hence it remains unaligned. Such phrase pairs are liable to be rejected according to the authors in [1]. But, this is somewhat rigid constraint. The authors in the paper [6] claimed that phrasal translations restricted to syntactic phrases only generate poor translations. The proposed framework uses a little flexible approach. The syntactic constraints like these are used as feature of the classifier in the proposed framework. Alignment of the boundary words (first or last) in a phrase is used to

ISSN: 2233-7857 IJFGCN

generate binary valued features as explained next. Source side syntactic features are defined as:

- SBS: if the source phrase begins with a syntactic sub-phrase, then SBS=1, otherwise, SBS=0.
- SES: if the source phrase ends with a syntactic sub-phrase, then SES=1, otherwise, SES=0.
- SFA: if the first word in the source phrase is aligned, then SFA=1, otherwise, SFA=0.
- SLA: if the last word in the source phrase is aligned, then SLA=1, otherwise, SLA=0.
- SOW: if the source phrase contains only one word, then SOW=1, otherwise, SOW=0.

Table 1 enlists the values of the source side syntactic features defined above for the phrase pair "does not: नहीं". Similar syntactic features are also applied on the target side phrases. All these syntactic features take care to retain only complete syntactic phrases and remaining phrases are pruned.

Table 1. Source Syntactic Feature Values for the Phrase Pair

Phrase pair	**does not ||| नहीं**
SBS	0
SES	0
SFA	0
SLA	1
SOW	0

Marker word information of English is also used as a feature of the classifier. According to Marker hypothesis, Marker words are a set of words which are required to structure the non-Marker words in the sentence to express the meaning in a natural language [4]. The syntactic structure of a sentence is marked at the surface level by the Marker words. Marker words form a closed set of words having certain part of speech tag such as quantifiers, determiners, prepositions etc. Consider following English sentence:

"Kiran played all day happily in her home."

In this sentence, the Marker words are *all* (quantifier), *her* (pronoun), *in* (preposition). If the Marker words are replaced by their respective category, the structure of the given sentence becomes:

"Kiran played ⟨quantifier⟩ day happily ⟨preposition⟩ ⟨pronoun⟩ home".

The Marker words shown in ⟨ ⟩ work as placeholder and can be replaced by any other words of same category to form other meaningful sentence. After replacing *all* by *one*, *in* by *at* and *her* by *his*, the newly generated sentence will be:

"Kiran played one day happily at his home".

The Marker words work as phrase boundary and the sentence can be easily segmented at Marker words [2][10]. This characteristic of Marker words can be used to find valid phrases hence is utilized as a feature for the classifier. This feature gets its binary value based on whether the phrase begins with the Marker word or not. Grammatical structure of *Hindi* is very much complex than English. Marker word

ISSN: 2233-7857 IJFGCN

sets of these languages are different. Hence, only English Marker word information but not the *Hindi* Marker word information is used in the proposed framework. The p-value defined in Fisher's exact test is also used as a feature for the classifier.

4. Experimentation

For experimentation three English to *Hindi* parallel datasets are used. One dataset is obtained from IIT, Mumbai. Second dataset is bilingual sentence pairs corpus developed in Tatoeba project. Third dataset is HindEnCorp 0.5 from ÚFAL. First, the obtained data is cleaned by removing noise, inconsistencies in punctuation symbols and misaligned sentences. After doing some pre-processing like tokenization and true casing using MOSES, some limited data is used for experimentation as many systems are trained and each system required huge time for training process. As objective of the research is to find effective group of classifier features that prune the phrase translation table while keeping the translation quality intact and not to develop an operational translation system, only limited data is used for experimentation. Two types of datasets are used for testing purpose. In-domain test dataset is part of same data on which the PBSMT systems are trained but it is never used for training and tuning purpose. Out-of-domain test dataset is not related to training data and it is entirely different than it in topic, genre, style etc. Table 2 shows the statistic of all datasets after all pre-processing tasks.

Table 2. Dataset Used for Experimentation

Dataset type	Sentence count		Unique word count		Number of tokens	
	English	Hindi	English	Hindi	English	Hindi
Training Set	60003	60003	52290	60997	685325	751013
Development dataset	5205	5205	12392	13700	61039	66949
Out-of-domain test dataset	250	250	1191	1342	2555	2772
In-domain test dataset	250	250	1056	1296	2459	2731

4.1. Experimental Setup

The language model component of the PBSMT system is developed using freely available language modeling toolkit SRILM. 3-gram language model is developed using available *Hindi* language monolingual corpus. Kneser-Ney smoothing is used to deal with zero count n-grams problem of language modeling. Lexicalized reordering model with phrase-monotonicity-forward-fe-allff setting of parameters found effective for producing quality English to *Hindi* translations. The parallel sentences are word aligned using freely available toolkit GIZA++. This tool applies an EM algorithm to find word alignments between parallel sentences. *Hindi* to English and English to *Hindi* word alignments are obtained using GIZA++ and then symmetrized with grow-diag-final-and heuristic strategy. From this word alignment phrases are extracted. For extracting phrases and developing PBSMT system, freely available MOSES toolkit (03 Feb. 2015, Version 3.0) is used. The beam size utilized for pruning is 200 and top 20 translation options are utilized to produce n-best list for each input sentence. After going through many experimental comparisons, it is decided to use SVM for the classification purpose. SVM is a large margin classifier i.e. it provides optimal margin gap between class separating hyperplanes. Widely used open source toolkit LIBSVM is used for classification. RBF kernel function is used after comparing the accuracy of classification with all the kernel functions. Various subsets of the phrase translation table as given in Figure 4 in section 3 are obtained by tracing the decoding of sentences in development dataset. All the sentences in the corpus are parsed to obtain syntactic features from the parse trees. Freely

ISSN: 2233-7857 IJFGCN

available LTRC Shallow Parser (IIITH, Hydrabad) is used for parsing in this research work.

5. Result Analysis

5.1. Comparison of Various Feature Combinations

Many SVM based classifiers are developed by taking various combinations of features and their performance is tested. PBSMT system parameters and kernel functions are also tuned differently as per the requirement to increase the translation accuracy for each feature combination. The features used are: lexicalized reordering probabilities (LRP), phrase and lexicalized phrase translation probabilities (PLPTP), target syntactic features (TSF), source syntactic features (SSF), source Marker word information (SMWI), Len_Ratio which refers to the length ratio of phrases, the statistical significance metric (P_Value), length of the source phrase (L_e), length of target phrase (L_h). As average sentence length of English and it's translation in *Hindi* are different, two ratios (L_h)/(L_e) and (L_e)/(L_h) are also considered. The classifier features are systematically grouped together to check their effect on the translation accuracy. First, a group containing translation probabilities, P_Value and Len_Ratio is constructed and its table reduction power and translation accuracy is obtained. Remaining features are then added one by one in this group. The distortion model feature (LRP) is added first and then the source syntactic and the target syntactic features are added one at a time and both at last. Source Marker word information feature is also added in this group at last. Table 3 compares reduction in the phrase translation table size and BLEU and TER scores corresponding to the particular combination of features.

Table 3. Percentage Reduction and Translation Quality for Different Feature Combinations

Combination of Features	Number of Phrases	Approx. % reduction in Size	In-domain Test Data		Out-of-domain Test Data	
			BLEU	TER	BLEU	TER
Baseline System	17158	-	47.0	44.8	23.7	67.2
PLPTP+Len_Ratio+P_Value	97814	43	46.9	45.0	23.3	67.4
PLPTP+LRP+Len_Ratio+P_Value	75438	56	46.8	45.1	23.2	67.5
SSF+TSF	85763	50	46.1	46.9	22.5	68.3
SSF+TSF+SMWI	92616	46	46.9	44.9	23.5	67.2
PLPTP+LRP+Len_Ratio+P_Value+	73722	57	46.8	45.1	23.1	67.6
PLPTP+LRP+Len_Ratio+P_Value+	72056	58	46.8	45.1	23.1	67.6
PLPTP+LRP+Len_Ratio+P_Value+	65213	62	46.8	45.1	23.1	67.6
PLPTP+LRP+Len_Ratio+P_Value+	72026	58	46.8	45.1	23.1	67.6
PLPTP+LRP+Len_Ratio+P_Value+	66937	61	46.9	44.9	23.0	67.7
PLPTP+LRP+Len_Ratio+P_Value+ SSF+TSF+SMWI	566136	67	46.90	45.04	23.01	67.71
PLPTP+LRP+(L_h)/(L_e)+(L_e)/(L_h)+	10636	38	46.9	44.9	23.4	67.3
PLPTP+LRP+(L_h)/(L_e)+(L_e)/(L_h)+	10291	40	46.9	44.9	23.5	67.2
PLPTP+LRP+(L_h)/(L_e)+(L_e)/(L_h)+	99549	42	46.9	45.0	23.5	67.2
PLPTP+LRP+(L_h)/(L_e)+(L_e)/(L_h)+ P_Value+SSF +TSF	892367	48	46.89	45.03	23.33	67.43
PLPTP+LRP+(L_h)/(L_e)+(L_e)/(L_h)+ P_Value+SSF +TSF+SMWI	754981	56	46.95	44.94	23.21	67.51

ISSN: 2233-7857 IJFGCN

The results show that, PLPTP, LRP, Len_Ratio, P_Value, SSF, TSF and SMWI features combination is effective for removing the phrase pairs and this combination can prune the translation table up to 67% of the original size. This combination restricts the low translation probabilities from being added into the translation lattice and get discarded without much affecting on the quality of translation. It is also observed that, whenever multiple options are available, the confusion is resolved in this classifier by taking decision in favor of those phrase pairs that satisfy syntactic phrase and Marker word information constraint. The syntactic features used in isolation without any other features support results in moderate classifier to prune the translation table. The classifier with the SSF and TSF combination only can remove approximately half of all the phrase pairs. BLEU and TER score of this PBSMT system gets declined as compared to other classifiers because the illegal syntactic phrases are allowed to get added in the translation model

5.2. Effect of Different Scale of Negative Data

For classification technique it is observed that, ratio of the scale of positive and negative examples is also having a large impact on the performance of the classifier. More negative training examples lead to removal of many phrases from translation model which degrades the translation quality. Therefore, judgmental balance must be maintained between the positive and negative examples and the translation quality. To study the effect of various scale ratios of positive and negative examples on translation quality and pruning rate, 1:1 to 1:4 different scale ratios of positive to negative examples respectively are used. Totally, 10000 positive samples and 10000 to 40000 negative samples with an increment of 10000 are selected. It is observed that, if the classifier is trained to be more sensitive for negative data then it removes more phrase pairs. But beyond the ratio of 1:2, the pruning rate converges at 67%. To conclude, to get maximum reduction while keeping comparable translation quality, negative training examples must be double of the positive examples.

5.3. Comparison of Various Methods

In Table 4, comparison of results with previous hard rule oriented methods is given to find the effectiveness of the proposed framework. As each phrase table pruning method is using different theoretical characteristics and technique, each pruning method shows different efficiency at the phrase table pruning task.

Table 4. Comparison with Existing Techniques

Pruning Technique		Number of Phrases	% Reduction in size	In-domain Test Data		Out-of-domain Test Data	
				BLEU	TER	BLEU	TER
Baseline System		1715868	-	47.08	44.84	23.71	67.21
TMS		754481	56	45.10	48.50	21.40	70.74
USF		1407211	18	44.90	49.10	21.20	70.79
Len_Ratio		1063338	38	39.90	59.90	19.70	79.80
Statistical Based Methods	Count Based	1235224	28	44.80	48.90	21.30	69.90
	Probability	1201607	30	44.95	46.90	22.30	69.70

ISSN: 2233-7857 IJFGCN

	Threshold	892356	48	45.70	47.60	22.25	68.60
	Histogram	840375	51	45.60	47.80	22.10	68.80
Fisher's Test (thr = α-ϵ)		669294	61	46.08	45.84	22.61	68.21
Relative Entropy		617412	64	46.20	45.50	22.71	68.10
Fisher's Test with Relative Entropy		600353	65	46.50	45.34	22.91	68.05
Proposed Framework in This Paper		566136	67	46.90	45.04	23.01	67.71

The results are compared using following factors .

5.3.1. Phrase Table Size

The effectiveness of pruning techniques is compared by measuring percentage reduction they achieve when calculated against original whole phrase table of baseline PBSMT system without any pruning. Percentage reduction achieved for all techniques along with proposed framework is given in Table 4. The proposed pruning framework reduces the phrase translation table by 67% from the original size.

5.3.2. Quality of Translation

The system generated translation that is having more BLEU score and small TER value is considered as the best translation. The baseline PBSMT system without any pruning produces the output with 47.08 BLEU score and 44.84 TER value for in-domain test data and 23.71 BLEU score and 67.21 TER value for out-of-domain test data. The PBSMT system pruned using proposed framework generates the output with 46.90 BLEU score and 45.04 TER value for in-domain test data and 23.01 BLEU score and 67.71 TER value for out-of-domain test data. However, it is observed that BLEU score is having a minor dip (0.18 & 0.70) in BLEU score and a minor increase (0.20 & 0.50) in TER value for in-domain and out-of-domain test data when the quality of translation is compared against BLEU and TER value of baseline system. Even though BLUE score is decreased and TER value is increased by a minor value, it gives two major advantages. The proposed framework reduces the size of the phrase table due to which memory and processing time requirement of the translation system gets decreased by a reasonable value. The minor decrease in BLUE score and increase in TER value can be compromised against these greater gains.

Len_Ratio produces the poorest translations for which BLEU and TER values are 39.90 and 59.90 respectively for in-domain test data and 19.70 and 79.80 respectively for out-of-domain test data. Fisher's test with relative entropy reduces the phrase table by 65% and also produces the quality translations for which BLEU and TER values are 46.50 and 45.34 respectively for in-domain test data and 22.91 and 68.05 respectively for out-of-domain test data. The proposed framework reduces the phrase table by 67% that is 2% more than Fisher's test with relative entropy technique. The proposed framework also produces comparatively better translations with 0.40 points more BLEU score and 0.30 points lower TER score for in-domain test data and 0.10 points more BLEU score and 0.34 points lower TER score for out-of-domain test data than Fisher's test with relative entropy.

5.3.3. Translation Time

The translation rate is measured in terms of number of sentences translated per unit time by the translation system. The baseline translation system without pruning shows 120 sentences per minute translation rate. With this rate, baseline system

ISSN: 2233-7857 IJFGCN

takes maximum of 2.1 minutes' translation time to completely translate all the sentences in the in-domain test dataset. Average translation rate of translation system pruned with the framework proposed in this paper is 171 sentences per minute. With this rate, the proposed framework takes maximum of 1.27 minutes' translation time for the same in-domain test dataset. This indicates that the proposed pruning framework significantly reduces the translation time by 28% as compared to translation time taken by baseline system. The baseline system has to process the huge phrase translation table data to complete the translations whereas proposed pruning framework based PBSMT system processes only a small part of the same phrase translation table to complete the translations. This indicates that the proposed framework optimizes the translation time due to lower computation load. This time optimization is really helpful for real time translation systems to produce translations at much faster rate.

Table 5 compares statistical details of phrases in the pruned phrase tables obtained for all pruning techniques. In original phrase table the average candidate options per distinct source phrase are over estimated. The classifier proposed in this paper pruned most of the options and retained only diverse translation option set for each source phrase. Another observation from the Table 5 is about lengths of source and target phrases. Basically, *Hindi* sentences obtained for English sentences may be longer because for a single English word many *Hindi* words may be required to express the meaning. Hence, phrase length also increases. However, as observed in ALDSP and ALTP columns in Table 5 for the pruned tables, well translated phrases found to be of equal length.

Table 5. Statistical Details of the Phrases in Pruned Tables

Pruning Technique	Number of Phrases	DSPN	ACPDSP	ALDSP	ALTP
Baseline System	1715868	549077	3,11	4.37	4.51
TMS	754481	410778	1.83	4.07	4.11
USF	1407211	548734	2.69	4.10	4.13
Len_Ratio	1063338	563834	2.79	4.11	4.17
Count Based	1235224	630066	2.63	4.21	4.27
Probability	1201607	624575	2.83	4.17	4.21
Threshold	892356	526428	2.45	4.13	4.24
Histogram	840375	521280	2.40	4.17	4.23
Fisher's Test	669294	390359	1.96	3.97	4.07
Relative Entropy	617412	372287	1.86	3.81	3.85
Fisher's Test with Relative Entropy	600353	347084	1.79	3.87	3.89
Proposed Framework in This Paper	566136	397551	1.35	3.85	3.90

DSPN: distinct source phrase numbers, ACPDSP: average candidate options per distinct source phrase, ALDSP: average length of distinct source phrases, ALTP: average length of target phrases

6. Conclusion and Future Scope

Small size of phrase translation table gives many advantages from storage and computation point of view. To conclude, the proposed framework exhibits following advantages over existing pruning methods. (1) It exploits and integrates many linguistic features in a classification technique due to which it effectively removes approximately 67% of the phrase pairs from the phrase table while keeping comparable quality of translation. (2) Only two factors need to be determined for the

ISSN: 2233-7857 IJFGCN

proposed classifier. First factor is the classifier training data and the second is classifier features.

In future, proposed framework will be explored for some other language pairs by using additional linguistic features available in those language pairs along with the characteristics used in this paper. Also, the proposed framework is not limited to pruning of a phrase based translation model only. It may be extended for any other models like hierarchical and syntax translation models. Quality of translation may be further improved by using similarity measurement techniques.

References

[1] H. Cao, A. Finch, and E. Sumita, "Syntactic Constraints on Phrase Extraction for Phrase-Based Machine Translation", Proceedings of the SSST-4, 4th Workshop on Syntax and Structure in Statist.Ttranslat., Beijing, China, **(2010)**, pp. 28–33.

[2] N. Chatterjee and S. Gupta, "Efficient Phrase Table Pruning for Hindi to English Machine Translation Through Syntactic and Marker-Based Filtering and Hybrid Similarity
Measurement", Natural Language Engineering, Cambridge University Press, **(2018)**, pp. 1-40, DOI: 10.1017/S1351324918000360.

[3] M. Eck, S. Vogel, and A. Waibel, "Translation Model Pruning Via Usage Statistics for Statistical Machine Translation", Proceedings of the HLT-NAACL (Short Papers), Rochester, NY, USA, **(2007)**, pp. 21–24, DOI:10.3115/1614108.1614114.

[4] T. Green, "The Necessity of Syntax Markers. Two Experiments with Artificial Languages",
Journal of Verbal Learning and Behavior, vol. 18, **(1979)**, pp. 481-96, https://doi.org/ 10.1016/S0022-5371(79)90264-0.

[5] J. H. Johnson, J. Martin, G. Foster, and R. Kuhn, "Improving Translation Quality by Discarding Most of the Phrase Table", Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL '07), **(2007)**, pp. 967–975.

[6] P. Koehn, F. J. Och, and D. Marcu, "Statistical Phrase-Based Translation", Proceedings of the Conference of the North American Chapter of the Association for Computational
Linguistics on Human Language Technology, Stroudsburg, Pa, USA, **(2003)**, pp. 48-54, DOI: 10.3115/1073445.1073462.

[7] K. Kavitha, L. Gomes, and G. Lopes, "Using SVMs for Filtering Translation Tables for
Parallel Corpora Alignment", Proceedings of the EPIA, **(2011)**, https://research.varia ncia.com/pubs/epia2011filter.pdf.

[8] W. Ling, J. Graca, I. Trancoso, and A. Black, "Entropy-Based Pruning for Phrase-Based Machine Translation", Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Jeju Island, Korea, **(2012)**, pp. 962–971, https://www.aclweb.org/anthology/D12-1088.pdf.

[9] T. Mei, Z. Yu, and Z. Chengqing, "Exploring Diverse Features for Statistical Machine Translation Model Pruning", IEEE/ACM Trans. on Audio, Speech and Lang. Proc., vol. 23, no. 11, **(2015)**, pp. 1847-1857, DOI: 1847 1857. 10.1109/TASLP.2015.2456413.

ISSN: 2233-7857 IJFGCN

[10] F. S´anchez-Mart´ınez and A. Way, "Marker-Based Filtering of Bilingual Phrase Pairs for SMT", Proceedings of the EAMT, the 13th Annual Meeting of the European Association for Machine Translation, Barcelona, Spain, **(2009)**.

[11] N. Tomeh, N. Cancedda, and M. Dymetman, "Complexity-Based Phrase-Table Filtering for Statistical Machine Translation", Proceedings of the MT Summit XII, Ottawa, Canada, **(2009)**, http://www.mt-archive.info/MTS-2009-Tomeh.pdf.

[12] N. Tomeh, M. Turchi, G. Wisinewski, A. Allauzen, and F. Yvon, "How Good are Your Phrases? Assessing Phrase Quality with Single Class Classification", Proceedings of the Int. Workshop on Spoken Lang. Translat., San Francisco, USA, **(2011)**, pp. 261–268.

[13] L. Wang, T. Nadi, X Guang, B. Alan, and T. Isabel, "Improving Relative Entropy Pruning Using Statistical Significance", Proceedings of the 25th International Conf. of Compu. Linguist. (Posters), Mumbai, India, **(2012)**, pp. 713-722, https://www.aclweb.org/anthology/C12 2070.pdf.

[14] L. S. Zettlemoyer and R. C. Moore, "Selective Phrase Pair Extraction for Improved Statistical Machine Translation", Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers, Association Computational Linguistic, Rochester, New York, **(2007)**, pp. 209-212, https://www.aclweb.org/anthology/N07-2053.pdf, DOI: 10.3115/1614108.1614161.

[15] R. Zens, D. Stanton, and P. Xu, "A Systematic Comparison of Phrase Table Pruning Techniques", Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Jeju Island , Korea,**(2012)**, pp. 972-983, https://www.aclweb.org/anthology/D12-1089.pdf.

Authors

Arun Babhulgaonkar pursed Bachelor of Engineering from Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, Maharashtra, India and Master of Technology from Dr. Babasaheb Ambedkar Technlogical University, Lonere, Maharashtra, India. He is currently pursuing Ph.D. and working as an assistant professor in Department of Information Technology at Dr. Babasaheb Ambedkar Technlogical University, Lonere, Maharashtra, India since 2004. He is a lifetime member of ISTE since 2008. He has published 06 research papers in reputed international journals and conferences including IEEE. His main research work focuses on Natural Language Processing, Machine Learning. He has 15 years of teaching experience.

Shefali P. Sonavane pursed Ph.D. in 2010 in Computer Science and Engineering at Walchand College of Engineering, Sangli (Government aided autonomous institute) affiliated to Shivaji University, Kolhapur, Maharashtra, India. With two years of industry experience, she opted teaching as a career profession. Her Ph.D. work is supported under young scientist, research scheme by Department of Science and Technology, New Delhi, India. Dr. Shefali received best teacher award in 2008 and is a member of many professional organizations. Currently, she is working as an associate professor in Department of Information Technology at Walchand College of Engineering, Sangli. She has received research funds from DST and AICTE for various technical projects promoting work in the area of Computer Vision and Information Security. She has a good number of

ISSN: 2233-7857 IJFGCN

publications in journals and participation in conferences with few IPR credentials at her account. She has extended her research interest further in the field of Machine Learning and Big Data. She is an active member towards the implementation of Outcome Based Education (OBE) in engineering with special efforts in revamping the teaching methodology and its assessment.

ISSN: 2233-7857 IJFGCN