

Improvement of the Results of Statistical Machine Translation System using Anusaaraka

Shubhamay Sen
School of Computer Engineering
KIIT University
India

Sriram Chaudhury
School of Computer Engineering
KIIT University
India

ABSTRACT

This paper describes an efficient experimental approach for the improvement of translation quality of phrase based statistical machine translation system by utilizing the insights of the rule based machine translation. As the most primitive step it is believed that appending large and accurately designed linguistic resources such as multiword bilingual dictionaries to the existing training corpus contributes a lot towards the enhancement of phrase alignment quality and phrase coverage of the Statistical Machine Translation (SMT) system. Further improvement in translation coverage can be achieved by improving the dictionary by introducing morph-syntactic word forms of the foreign language words instead of simple root word forms, and its corresponding translations in native language. As in real time testing scenario, the test corpus may possess different morphological extensions of the root word which is not covered by standard dictionaries. As a matter of fact addition of such dictionaries to the corpus enriches it and provides a solution to the improper translations previously generated due to occurrences of morph-syntactic extensions instead of the root word form. As the proposed approach towards further improvement, the intelligence of Anusaaraka and huge computational ability of SMT is integrated to achieve better translations. Anusaaraka is a machine translation system based on Panini's Astadhyayi grammatical rules and an expert when the English-Hindi phrase alignment is concerned. It does it by comparing its output translation with the accurate manual translation and extracting out the best possible option. The bi-lingual phrase pairs thus obtained are highly accurate and when appended to the training corpus of statistical machine translation system results as better phrase alignment structure, hence better translation quality.

General Terms

Phrase based statistical machine translation, Statistical machine translation, Machine translation, Rule based machine translation.

Keywords

Statistical machine translation, Bilingual dictionary, Morphological dictionary, Anusaaraka, Phrasal, Phrase alignment.

1. INTRODUCTION

A major portion of the information available in electronic media is in English. But when the common men are concerned, a very small proportion of world's population knows English. Inability to access the text in a foreign language creates the language barrier and leads to digital-divide. So machine translation can be an effective tool for common men to understand the information encoded in foreign languages and to communicate with others, thus promoting the globalization of information system. In this prospect advancements in machine translation are utmost

desirable. Machine translation is an automated system that analyzes text from source natural language and translates it to target natural language conveying the exact meaning of the input text and producing fluent text in the output language, preferably without human intervention. Many researchers have successfully tried different approaches in this domain over 60 years and the translation results thus received are quite appreciable. But it is an unavoidable fact that Fully-automatic General purpose High quality Machine Translation systems (FGHMT) are hard to develop and hence till date no such machine translation system exists that qualifies as an FGHMT. Especially when a phrase based statistical machine translation system for English- Hindi pair of language is concerned, the results are not quite convincing. One of the possible reasons is the languages differ a lot in the sentential structure as Hindi in comparison to English is a free word-order language, where the noun phrases can come in any order followed by the verb phrase. Hence an approach for improving the performance of the statistical phrase based machine translation system (*Phrasal* machine translation system) for English-Hindi language pair by utilizing the phrase level alignments extracted from the Anusaaraka machine translation system is proposed.

Anusaaraka [1][3][6][7][15] is a language access cum machine translation system based on Panini's Astadhyayi grammar rules and believes on the concept of sharing of load between men and machine. According to this approach a text of any language contains only a part of the actual information the speaker wants to convey. Rest the reader itself perceives by analyzing it with the help of world knowledge, context knowledge, language conventions and common sense. As machines do not have any common sense or context specific knowledge, it's capability of interpreting a general text with desired level of accuracy is very low. A simple text in this context is *love of mother*. In this case the machine faces a common difficulty in identifying the subject (कर्ता) and object (कर्म). So during translation to target language text the possible interpretations may be "*Mother's love*" or "*Love for mother*", i.e. "मां का प्यार." else "माँ के लिए प्यार". Human beings having context knowledge can easily resolve this but the machine can't. Same in case of the sentence; *Tom saw a tourist in the park with a telescope*. Can be interpreted as any of the following i.e. '*saw with a telescope*', '*tourist with a telescope*' or '*park with a telescope*'. When the common sense is concerned, two famous examples in this context are; 1) The mother with babies under four... and 2) The mother with babies under forty.... While translating it the machine may produce silly results such as "चार के तहत बच्चों के साथ मां..." and "चालीस के तहत बच्चों के साथ मां..." due to the lack of common sense. Some idiomatic expressions may also be the reason of such complicacy. An example in this context is, *He kicked the ball* and *He kicked the bucket*. In the first case it really means the act of kicking the ball, where as the next sentence being an

idiomatic expression, the phrase *kicked the bucket*, refers to the act of *dying* not to the literal meaning of kicking a bucket. But machines lacking such intelligence may produce erroneous results i.e. “वह गेंद को लात मारी.” and “वह बाल्टी लात मारी.”. The proposed approach tries to resolve such issues and to improve the phrase alignment structure (English-Hindi language pair) of the statistical model in order to synthesize appropriate output Hindi *Padas* (phrases).

2. STATISTICAL MACHINE TRANSLATION SYSTEM: AN OVERVIEW

Statistical Machine Translation is the one of the most excelled machine translation approaches based on the concept of statistical models, whose parameter are estimated by analyzing huge bilingual parallel text corpora. In phrase-based translation systems like Moses, Phrasal etc the units of translation are phrases, might be of different length, which are statistically extracted out of bi-lingual parallel corpus and most preferably not the linguistic ones. The major advantages of statistical machine translation over other approaches are standardization of approach, less human effort and natural translation. But as an unavoidable part there exist some hindrances such as need of large bilingual sentence aligned parallel corpora, which may not be present and development of the same may involve too much time as well as cost. It may also result as some irrelevant translations as it is a complete probabilistic approach. Translation performance may also degrade for the language pair those belong to different language family having entirely different sentence ordering structure. Further insufficiency of training data may lead to unavailable word forms or phrases, resulting as un-translated phrases in the final translated text.

2.1 Phrase Based SMT Architecture

The development of phrase based SMT (Phrasal) model undergoes three distinct steps as word alignment, phrase pair extraction and scoring of phrase pairs.

2.1.1. Word alignment [11][12][13][14][18]

Word alignment is the process of mapping words of one language to that of the other of the language pair and vice-versa. The alignment function “*a*” is defined as $a: i \rightarrow j$. Where i = Target language word at position ‘*i*’ and j = Source language word at position ‘*j*’.

2.1.2. Phrase pair extraction [11][12][13][14]

All possible statistical phrase pairs are extracted out of the bi-lingual training corpus. Phrase pair (\bar{e}, \bar{f}) is consistent with an alignment “*A*”, if all words f_1, \dots, f_n in f that have alignment points in A have these with words e_1, \dots, e_n in e and vice versa.

2.1.3. Scoring of phrase pairs [11][12][13][14]

Extracted phrase pairs are scored by calculating the relative frequency of the pairs using Bayes’ Rules.

$$e_{best} = \operatorname{argmax}_e P(e/f) = \operatorname{argmax}_e P(f/e) P_{LM}(e)$$

Decomposing translation model,

$$p(\bar{f}_1^l / \bar{e}_1^l) = \prod_{i=1}^l \Phi(\bar{f}_i, \bar{e}_i) d(\operatorname{start}_i - \operatorname{end}_{i-1} - 1)$$

Where, $P(f|e)$ = Phrase translation model, $P_{LM}(e)$ = Language model, Φ = Phrase translation probability, d = Re-

ordering probability, l = Number of phrases in the source language text. Extending the language model;

$$e_{best} = \operatorname{argmax}_e \prod_{i=1}^l \Phi(\bar{f}_i, \bar{e}_i) d(\operatorname{start}_i - \operatorname{end}_{i-1} - 1) \prod_{i=1}^{|\bar{e}|} p_{LM}(e_i / e_1 \dots e_{i-1})$$

2.1.3.1. Phrase translation model

The probability of phrase translation for any given phrase pair is defined as the ratio between the total numbers of times the given phrase pair is applicable to the total number of times the phrase pair having the source language text as its first phrase appears in the corpus.

$$\Phi(\bar{f}, \bar{e}) = \frac{\operatorname{count}(\bar{e}, \bar{f})}{\sum_{\bar{f}_i} \operatorname{count}(\bar{e}, \bar{f}_i)}$$

Where, \bar{f}_i = any source language phrase to be translated, \bar{e}_i = Corresponding target language phrase from phrase pair, $\Phi(\bar{f}_i | \bar{e}_i)$ = Score from phrase translation table.

2.1.3.2. Reordering model

Distance based re-ordering scheme is adopted for re-ordering the target language phrases thus generated after phrase translation, as the two languages may follow different structural layout. The reordering model can be defined as $d(\operatorname{start}_i - \operatorname{end}_{i-1} - 1)$. Where, start_i = Starting index of the Target Language (TL) phrase in Source Language (SL), end_{i-1} = End index of the previous phrase (in TL) in SL.

2.1.3.3. Language model [9][11][12]

Defines how likely a string of target language words forms a good sentence of that language. It adopts n-Gram approach to determine the likelihood. It’s a probabilistic approach that defines, higher the P_{LM} value better the target language text. The target language sentence is denoted by “*W*”. Where “*W*” is a string of words (*Markov Chain*) as $w_1, w_2 \dots w_n$. For n-Gram

$$p(w_1, w_2, w_3, \dots, w_n) = p(w_1) p(w_2/w_1) p(w_3/w_1, w_2) \dots p(w_n/w_1, w_2, \dots, w_{n-1})$$

The Maximum Likelihood $P(w_i)$ for any word w_i is estimated by taking into consideration the last $n - 1$ words depending upon the n-Gram model used. For a bi-Gram model,

$$\operatorname{Max. Likelihood} = P(w_2/w_1) = \frac{\operatorname{count}(w_1, w_2)}{\operatorname{count}(w_1)}$$

2.1.3.4. Perplexity [9]

Perplexity is a convenient measure for comparing the language models. In-case the test data is too long then W_1^n will be assigned with quite a small value as the n-gram conditional probabilities themselves are very small. So $H(W)$ will be even smaller hence inconvenient to access. Hence Perplexity is calculated to obtain a relatively larger number which is easier to analyze.

$$\operatorname{Perplexity}(W) = 2^{H(W)}$$

Where, $H(W) = \frac{1}{n} \log_p(W_1^n)$ is defined as the Cross Entropy. “ n ” is the number of words in the test data. Dividing by “ n ” helps to normalize things. A good model will have a relatively large $P(e)$ and a relatively small perplexity.

Adding weights $\lambda_\phi, \lambda_d, \lambda_{LM}$ to the model,

$$e_{best} = \underset{e}{\operatorname{argmax}} \prod_{i=1}^I \Phi(\bar{f}_i, \bar{e}_i)^{\lambda_\phi} d(\text{start}_i - \text{end}_{i-1} - 1)^{\lambda_d} \prod_{i=1}^{|e|} p_{LM}(e_i/e_1 \dots e_{i-1})^{\lambda_{LM}}$$

The parameter λ defines the strength of the lexical weight p_w . Good values for this parameter are around 0.25.

2.1.4. Decoder [11][12][13]

The decoder implements a beam search technique. Provided an input string of words, a number of phrase translations are possible. The search begins in an initial state where no source language input words are translated and no target language output words are generated. Starting from the initial state, the first source language word is translated and is marked (*) as translated. The decoder proceeds from left to right. Further new states can be explored by translating rest of the uncovered words. The current cost of the new state is the product of the cost of the original state and the translation, distortion and language model costs of the next translation. The hypotheses those cover all foreign words with lowest possible cost (highest probability) are selected as best translation.

3. PREVIOUS WORK

Vogel and Monson [21] first introduced the idea of augmenting manual dictionaries to the training corpus to improve the performance of SMT systems by improving the training process. But dictionaries only contain the root forms of the words. So the translations generated may undergo some drawbacks like producing singular word forms where plurals needed in case of nouns or a completely wrong word-form where some inflections need to be generated in case of verbs. So they emphasized on introducing morphological variations of the root word forms in the appended dictionary. Adding new entries to the dictionary is a two step approach; i.e. morphological inflections of the word are generated by identifying the Part Of Speech for the English translation that may contain several tags. As for nouns: singular/plural or entries with definite and indefinite determiner and for Verbs: -ing, -ed, -s, un-, im- etc forms are generated. A large monolingual corpus is used to filter out the entries that don't appear in the corpus and the new entries are not added to lexicon. For assigning probabilities to the dictionary entries training corpus (bilingual-sentence aligned) is used. If the dictionaries entries are found in the corpus corresponding probabilities are assigned else default small probabilities are allocated.

Nießen [16] also supported the idea of adding conventional bilingual dictionaries as additional information resources to properly train the model parameters. But the major disadvantage of the above process is the absence of context which is available in case of full bilingual corpora. The word may have different readings depending on the context. As the word *bank* may have several meanings like a financial institution, river bank, turn on a road, a group of articles etc. based on the context. In normal bilingual corpora, the words can often be disambiguated by taking into account the sentence context in which they occur. In this work a technique for aligning corresponding readings of the word in conventional dictionaries containing pairs of fully inflected word forms is proposed. The approach utilizes the information extracted from one language side to solve the category

ambiguity of the corresponding entry in the other language side. The dictionaries thus obtained are best suited for enhancing the performance of machine translation. The performance is further improved when used along with the hierarchical lexicon models.

Burch et al. [4] prescribed that word level alignment of parallel corpus reduces the AER¹ and enhances the BLEU² [20] score (Performance) as comparison to sentence level alignment. They also revealed that if word level alignments are implemented in the training corpus with integrated dictionary it further reduces AER and enhances the performance.

Och [17] also provided experimental proofs revealing significant decrease in alignment error rate and improvement in translation quality when training on data with word-level alignments. Brown et.al developed a multinomial model for constructing a dictionary. This model uses a parameter called effective multiplicity $\mu(e, f)$ to add only eligible entries of dictionary to the training corpus and reject the unfit ones.

$$\mu(e, f) = \frac{\lambda(e) \cdot p(f/e)}{1 - e^{\lambda(e) \cdot p(f/e)}}$$

Where $\lambda(e)$ defines the size of the sample.

Och and Ney [18] suggested setting effective multiplicity with a large value ($\mu^+ \gg 1$) for a dictionary entry if the two words co-occur else with a lower value. So only frequent co-occurring entries are kept, rest are allocated with very small effective multiplicity value and hence suppressed. This is done to avoid out of domain words.

4. PROPOSED WORK

This entire experimental research is based on *Phrasal* [5]. *Phrasal* is an open source machine translation system basically developed for machine translation between European languages. First part of this work (Model1) defines the basic implementation of *Phrasal* SMT system for developing a machine translation model for English-Hindi language pair. That forms the base line system. The model is trained over the sentence aligned parallel bi-lingual (English-Hindi) corpus only. No additional resources are appended to the training corpus. The next consecutive parts describe several previously proposed approaches (influenced by some existing ideas for other languages) for the improvement of the performance of *Phrasal* English-Hindi Statistical Machine Translation system. Model2 explains a previously proposed approach for the improvement of the training process of the model, i.e. by implementing standard bilingual dictionaries to the existing training corpus. Model3 extends this idea. Model4 explains further improvement of the training corpus by introducing morpho-syntactic word forms (inflections) possible for the existing root words of the foreign language in the bilingual dictionary and thus enriching the dictionary contents. The last part (Model5) describes the proposed idea of integrating aligned multi-word bilingual phrase extracts from *Anusaaraka* machine translation system to the existing training corpus. For convenience of processing all of the Hindi text is represented in WX notations [2]. WX notation is a standard notation used to represent the Hindi text in Roman script. All the processing is done in WX itself. Converters are used to convert the utf-8 Hindi text to WX before processing by SMT and to do the reverse after generation of the translation results. *Anusaaraka*

¹ AER- Alignment Error Rate

² BLEU – Bi- Lingual Evaluation Understudy

is embedded with such converters. The experiment proceeds through the following five models.

4.1. Model 1

This base model uses only the bi-lingual parallel corpus to train the model as well as to develop and test it. The parallel corpus used to train the model consists of 12,868 English sentences, randomly chosen from different domains and its appropriate translations in Hindi. The translations are done by experts; so have a very higher rate of accuracy. After training the model on the above corpus it is tested on a test data set of 400 sentences. The translated Hindi texts thus generated are subjected to evaluation by BLEU [20] and Human evaluation methodology and is presented in table-2 and table-3 respectively. One of the basic drawbacks of this model is the appearance of out of domain words in the test corpus during real time testing. As these words were not encountered during training process, so may lead to either omission of that word or reappearance of it in the final translated text as it is in the input text. Addition of large standard dictionaries leads to avoidance of such instances, as dictionaries possess words from several possible domains.

4.2. Model 2

Conventional bilingual dictionaries are well designed sources of aligned phrase pairs which can contribute towards the proper phrase alignment structure of the Phrasal SMT model. So as a previously proposed approach [16][21] towards improvement of English-Hindi Phrasal SMT System, properly designed bilingual dictionaries are appended to the existing bilingual training corpus. Sabdanjali: an English-Hindi dictionary containing almost 30,500 phrase pairs is adopted as the standard conventional dictionaries to be implemented to the corpus, as an experimental approach to prove the concept. The tokenized monolingual parts of the dictionary are appended to the corresponding parts of the training corpus. The Phrasal SMT is now trained over the above data. The phrase alignment structure and bi-lingual phrase pairs thus obtained after the training phase shows obvious improvement in the alignment quality as well as coverage. The Phrasal MT model thus developed is tested over a test corpus of 400 foreign language (English) sentences, and the native language (Hindi) translations thus obtained is compared with the manual reference translations of that test data. Final scores (BLEU score & Human Evaluation score) thus obtained is presented in table-2 and table-3.

4.3. Model 3

The previous experiment is now extended in Model3 by appending a comparatively larger multi-word English-Hindi dictionary extracted out of Anusaaraka MT system containing almost 97,000 phrase pairs. This dictionary is a multi-word phrase aligned corpus that consists of one or more words at both the language side. The entire process is repeated again for training the SMT. Developed model is again tested over the same test corpus. Final results are represented in table-2 and table-3.

4.4. Model 4

The above experiments reveal the generation of some inappropriate translations like producing singular word forms where plurals needed in case of nouns or a completely wrong word-form where some inflections need to be generated in case of verbs. Main cause behind this is that standard dictionaries only contain the root word forms in source language and its corresponding translation in target language. But in the real time test scenario the system may encounter

several morph-syntactic forms [21] of a source language word other than the root word itself in the test corpus. Those words may remain unhandled, wrongly aligned or even may be dropped. So as the third step of improvement of the training model the dictionary is improvised by introducing morph-syntactic extensions. An English-Hindi bilingual dictionary consists of different possible word forms on the basis of pos tag status like noun, verb, adjective, proper nouns, prepositions etc. The words tagged as nouns can further be categorized on the basis of number i.e., singular and plural and verbs on the basis of tense, aspect, person, gender and number. Apertium [8] open source machine translation system is used for the generation of different such forms of the English words from a standard monolingual word set. A sample extraction is as below;

The root word “play” can have the following extensions depending on the morph-syntactic tags.

```
play: play<vblex><inf>
play: play<vblex><pres>
plays: play<vblex><pri><p3><sg>
played: play<vblex><past>
played: play<vblex><pp>
playing: play<vblex><ger>
playing: play<vblex><pprs>
playing: play<vblex><subs>
```

The enriched monolingual English dictionary thus obtained is further filtered by referring to the Oxford- British National Corpus (BNC) to extract out only those words which are frequent in day to day English vocabulary and to filter out those which hardly appear in general English. This can be stated as a quality enhancement step that enhances the corpus quality by keeping the size within a reasonable extent. The final dictionary thus obtained consists of almost 50,000 words (root words and their morphological extensions). The corresponding Hindi translations are then manually generated by taking as reference the translation output form Anusaaraka Machine Translation system. Further manual verification ensures proper translation as Hindi is language where translations are greatly influenced by gender and number factors. The morph-syntactic dictionary thus generated can be treated as a phrase aligned corpus where for a phrase in source language (English) consisting of a single word is aligned with the phrase of target language (Hindi) having one or more words. As the next step the dictionary is appended to the parallel training corpus and the Phrasal SMT is trained over that. The translated output Hindi text thus generated is now evaluated by BLEU and Human evaluation system.

4.5. Model 5

The Model5 describes the idea of integrating the phrase extracts of Anusaaraka machine translation system with SMT. Anusaaraka [1][3][6][7][15] is an English-Hindi machine translation system based on Panini's Ashtadhyayi (Grammar rules). The Anusaaraka researches reveal that for machine translation to be successful there needed a fusion between the Intelligence of human being and computational ability of the machine. Researches in this field also reveal that the phrase alignments generated by Anusaaraka machine translation system is much more appropriate than those of the statistical machine translation systems' when the English-Hindi language pair is concerned. Hence the idea of integrating the Anusaaraka machine translation system's linguistic ability with the computational power of statistical machine

translation system (Phrasal) is proposed. This experimental approach is as follows. Practically a large parallel bi-lingual sentence aligned corpus is needed for implementation in Anusaaraka MT system. But due to the absence of such corpus, a small replica corpus of size 2,000 aligned sentences is collected. As the latest feature of Anusaaraka MT system, it automatically yields the aligned phrase pairs for the above corpus when executed. It does it by comparing the phrase translations generated from Anusaaraka itself with the expert manual translations and the English-Hindi phrase alignments are further improved by choosing the better alternative. The output alignment phrase pairs are better aligned than those of the SMT but consist of few inappropriate entries that need to be cleaned manually as this is a new feature of Anusaaraka and yet not completely grown up. This requires a lot of effort and is the most time consuming part of the preprocessing. The refined phrase pairs are now appended with basic training corpus of the model, which enriches the models (Model5) training corpus quality. Thus aligned phrase pairs obtained after training the system on the current improved corpus are better as compared to the base line system. The model is now tested for the same test corpus of 400 English sentences. The output translations obtained are subjected to both BLEU and Human evaluation. Final scores thus received are represented in table-2 and table-3 respectively.

5. RESULTS AND PERFORMANCE ANALYSIS

5.1. Results

The experiment uses English-Hindi parallel bi-lingual sentence as well as phrase aligned corpus. Size of the corpus in number of sentences used for training, development and testing of statistical models are shown in table 1. The first model consists of parallel sentence aligned English-Hindi corpus only and the later four models are trained upon the same corpus with appended word else phrase aligned dictionaries. All the models are tested on the same test corpus. The corpus details sums of both the parallel corpus and appended dictionaries.

Table 1. Corpus details:

	Training	Development	Test
Model1	12868	600	468
Model2	43357	600	400
Model3	109505	600	400
Model4	61253	600	400
Model5	52013	600	400

A comparison between the sample translation outputs generated by the above five models for a given source language (English) text is shown below.

English: Jaipur, popularly known as the pink city, is the capital of Rajasthan state, India.

Hindi Translations:

Model1 गुलाबी नगरी के रूप में प्रसिद्ध जयपुर की राजधानी राजस्थान राज्य , है .

Model2 गुलाबी नगरी के रूप में प्रसिद्ध जयपुर है जयपुर राजस्थान राज्य , भारत की राजधानी .

Model3 गुलाबी नगरी के रूप में प्रसिद्ध जयपुर की राजधानी राजस्थान राज्य .

Model4 गुलाबी नगरी के रूप में प्रसिद्ध है जयपुर राजस्थान राज्य , भारत की राजधानी .

Model5 गुलाबी नगरी के रूप में प्रसिद्ध जयपुर है राजस्थान राज्य , भारत की राजधानी .

The above example translations define the successive improvement of the generated outputs from the five models. The output Hindi text from Model1 consists of a missing translation for *india*. The dropping of such words, as a result of improper phrase alignment of the model leads to inefficiency. There is also an improper ordering of phrases in the second part of the translated Hindi text, i.e. की राजधानी राजस्थान राज्य, है instead of राजस्थान राज्य की राजधानी है. The source text (English) wants to convey the information i.e. Jaipur is the capital of Rajasthan state of India but the actual Hindi translation conveys as Rajasthan is the capital of Jaipur that leads to an entirely different meaning. In the second translated text (Output of Model2) the model generates more than one instance of translation (जयपुर है जयपुर) for *jaipur*. Improper ordering of phrases is also encountered in the translated text of Model3. More proper translation will be “जयपुर राजस्थान राज्य की राजधानी” instead of “जयपुर की राजधानी राजस्थान राज्य”. The translations for *is* and *india* are also missing. The Model4 leads to quite an accurate translation except improper placement of the word है. Whereas the translated Hindi text obtained from Model5 (proposed model) is reasonably accurate with elimination of all the above drawbacks.

5.2. Performance Analysis

Two well-defined techniques are adopted to evaluate the developed models' performance i.e. BLEU scoring technique and Human Evaluation method. The results obtained are as below.

5.2.1. BLEU scoring [20]

BLEU metric is used to evaluate the translation quality of the final text generated from the models. It's an automatic system evaluation technique where the quality of the translated text is measured by comparing it against several manual reference translations. The BLEU metric ranges from 0 to 1 (1 for the best and 0 for the worst). A comparative study of the BLEU scores achieved is shown in table 2. Model1 denotes the SMT model trained over basic corpus only with no appended dictionaries. It forms the base-line system.

Table 2. BLEU scores

Models	BLEU
Model1	0.352
Model2	0.382
Model3	0.383
Model4	0.380
Model5	0.372

5.2.2. Human evaluation

Human evaluation methodology is adopted to measure the performance of the above models manually. The BLEU scoring technique though efficient, but does not guarantee the improvement of the models' translation quality. For evaluating the above models 40 output test translations are

randomly chosen from each of the five models' translated output test data along with its corresponding English text. The translations are now subjected to evaluation by five different judges. Most of them are native speakers of the target language (Hindi) but none of them is a professional translator. This criterion is maintained in order to mimic the real time scenario where the users of this system are common people not experts. Each of the evaluator marks the translation out of five (5 for best, 1 for worst and so on). After evaluation is completed the average of marks allocated by all the five judges against each of the translations (by a model) for a given English sentence is calculated. Now the overall mark obtained by a particular model for all translated sentences can be calculated by taking the average of all marks obtained by that particular model for all translations. A comparative study of the evaluation scores of all the five models is shown in Table 3. The scores denote the mean scores obtained by the above five models against all the 40 randomly chosen test translations. The next row denotes the % scores obtained by the models. A graph can be plotted from this statistical outcome as shown in Fig 1. The graph thus plotted shows a natural improvement of the system's performance over the baseline system.

Table 3. Human evaluation scores

	M1	M2	M3	M4	M5
Scores (Out of 5)	3.390	3.461	3.335	3.500	4.085
% Score	67.8	69.22	66.7	70	81.7

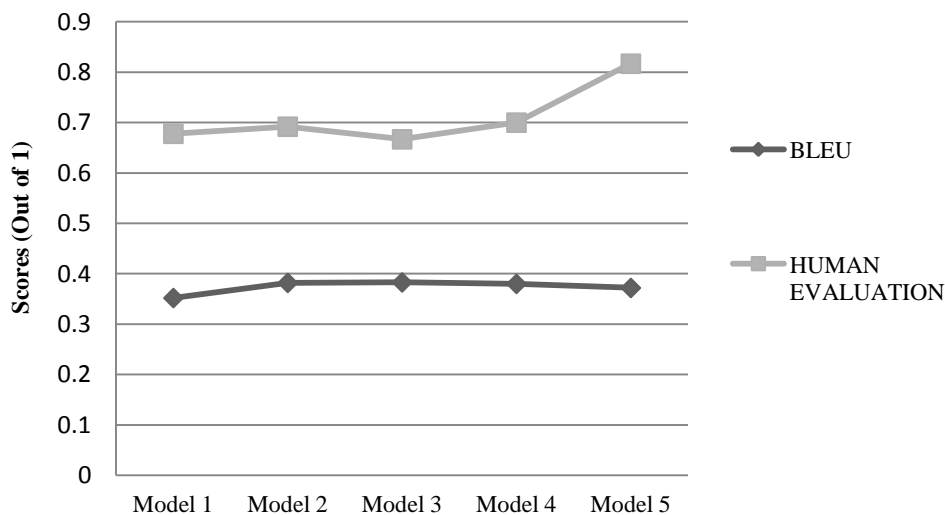


Figure 1: Performance graph

The work also proposes the development possibility of large bi-lingual phrase aligned dictionary for the same language pair by using alignments of Anusaaraka MT system, which would be indeed an efficient resource for phrase based machine translations. The Anusaaraka phrase alignment process is still in evolving stage and continuously improving. Hence this field promises a scope for future research, where the linguistic ability of Anusaaraka and computability of SMT can be integrated to achieve the optimum possible accuracy in translation. Further in all the above experiments that have

Although the graph seems to be a bit declining for the proposed model (Model 5) when the BLEU scores are concerned, it should be considered that higher BLEU score does not guarantee better translation. Further it is an experimental implementation to prove the idea; hence the training corpus used for the model is quite small. But being a statistical model the system would definitely perform well when trained with sufficiently large corpus as although declining, the proposed model scores better than the base line system (Model 1) in both the modes of evaluation. A possible source of some wrong translations is some improper phrase alignments generated from Anusaaraka machine translation system as the phrase alignment extraction feature is a recent addition to Anusaaraka and is in improvement stage. But better performance is desirable in recent future. The model also provides promising response when the Human evaluation is concerned as the graph shows a sharp growth. In real time scenario this model can be implemented by extracting a huge aligned phrase set from Anusaaraka.

6. CONCLUSION AND FUTURE RESEARCH

As basic initiative the work described above tried to improve the translation quality of Phrasal statistical machine translation system for English–Hindi language pair, by improving the quality of training corpus by implementing some existing and some entirely new ideas and received appreciable outcome by both the means of evaluation available. In this research some efficient resources like a morphological English-Hindi dictionary of 48,385 entries are also developed.

been performed the SMT models have been trained on quite a small corpus. But as a well proven fact the larger the size of training corpus the better the system performs. Hence all the above ideas can be implemented on large corpora, to make this idea use worthy in real time situations.

7. ACKNOWLEDGMENTS

First and foremost, the authors would like to express their deepest gratitude to the KIIT University, Bhubaneswar for facilitating the research with appropriate infrastructure. They

are grateful to Stanford NLP group for the Phrasal machine translation tool which forms the base of the research. They would also like to thank IIIT-Hyderabad & IIT-Bombay, for Anusaaraka machine translation system and for useful linguistic resources. They thank Apertium group for Apertium shallow parser and to all those who directly or indirectly contributed to the successful completion of this research.

8. REFERENCES

- [1] Anusaaraka, 2013. <http://anusaaraka.iiit.ac.in/>.
- [2] Bharati, Chaitanya, Sangal, 2000. Natural Language Processing: A Paninian Perspective, pp. 193. <http://anusaaraka.iiit.ac.in/node/65>
- [3] Bharati, Kulkarni, 2009. Anusaaraka: An Accessor cum Machine Translator. At: First Workshop on Free Rule Based MT, Alacante, Spain, 2nd November. Available FTP: http://sanskrit.uohyd.ernet.in/faculty/amba/PUBLICATIONS/presentation_spain.pdf.
- [4] Burch, Talbot, Osborne, 2004. Statistical Machine Translation with Word- and Sentence-Aligned Parallel Corpora. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL '04), Article No. 175.
- [5] Cer, Galley, Jurafsky, Manning, 2010. Phrasal: A Toolkit for Statistical Machine Translation with Facilities for Extraction and Incorporation of Arbitrary Model Features. In: Proceedings of NAACL Demo Session, Los Angeles, USA.
- [6] Chaudhury, Rao, Sharma, 2010a. Anusaaraka: An Expert system based MT System. In: Proceedings of IEEE conference on Natural language processing and knowledge management (IEEE-NLP KE 2010), Beijing, China.
- [7] Chaudhury, Sharma, Kulkarni, 2010b. Anusaaraka: An Approach to Machine Translation. In: Proceedings of the International Conference on "Language, Society and Culture in Asian Context". MSU, Thailand, 6-7 January.
- [8] Forcada, Bonev, Rojas, Ortiz, Sa´nchez, Mart´inez, Oller, Montava, Tyers, 2009. Documentation of the Open-Source Shallow-Transfer Machine Translation platform *Apertium*. In: Technical report, Departament de Llenguatges i Sistemes Inform`atics, Universitat d'Alacant, Alicante, Spain, 10th March.
- [9] Knight, 1999. A Statistical MT Tutorial Workbook, In: JHU summer workshop, 30th April. Available FTP: www.isi.edu/natural-language/mt/wkbk.rtf.
- [10] Knight, Koehn, 2003. What's New in Statistical Machine Translation, Information Sciences Institute, University of Southern California. Available FTP: homepages.inf.ed.ac.uk/pkoehn/publications/tutorial2003.pdf.
- [11] Koehn, 2013a. Tutorial on Statistical MT. <http://www.statmt.org/book/>.
- [12] Koehn, 2013b. Moses: Statistical Machine Translation System User Manual and Code Guide. University of Edinburgh. 26th August. Available FTP: <http://www.statmt.org/ Moses/manual/manual.pdf>.
- [13] Koehn, Hoang, Birch, Burch, Federico, Bertoldi, Cowan, Shen, Moran, Zens, Dyer, Bojar, Constantin, Herbst, 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In: Proceedings of the ACL 2007 Demo and Poster Sessions, Prague, June, pp. 177–180.
- [14] Koehn, Och, Marcu, 2003. Statistical Phrase-Based Translation. In: Proceedings of HLT-NAACL, Edmonton, Canada, May-June, Main Papers, pp. 48-54.
- [15] Kulkarni, 2003. Design and Architecture of `Anusaaraka'- An Approach to Machine Translation. In: Satyam Technical Review, vol 1, Q4, pp. 57-64, April.
- [16] Nießen, 2002. Improving Statistical Machine Translation using Morph-syntactic Information. In: Ph.D. thesis, Department of Computer Science, RWTH Aachen University, Aachen, Germany, December.
- [17] Och, 2002. Statistical Machine Translation: From Single-Word Models to Alignment Templates. In: Ph.D. Dissertation, Department of Computer Science, RWTH Aachen University, Aachen, Germany, October.
- [18] Och, Ney, 2000. A Comparison of Alignment Models for Statistical Machine Translation. In: Proceedings of the 18th conference on Computational linguistics (COLING '00), Saarbrücken, Germany, July, pp. 1086-1090.
- [19] Och, Ney, 2004. The Alignment Template Approach to Statistical Machine Translation. In: Computational Linguistics, Volume 30, Number 4, June.
- [20] Papineni, Roukos, Ward, Zhu, 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July, pp. 311-318.
- [21] Vogel, Monson, 2003. Augmenting Manual Dictionaries for Statistical Machine Translation Systems. In: Proceedings of Language Resources and Evaluation (LREC), pp. 1593-1596.