Handling verb phrase morphology in highly inflected Indian languages for Machine Translation

**Ankur Gandhe, Rashmi Gangadharaiah, Karthik Visweswariah,
and Ananthakrishnan Ramanathan**
IBM Research, India
{ankugand,rashgang,v-karthik,aramana2}@in.ibm.com

Abstract

The phrase based systems for machine translation are limited by the phrases that they see during the training. For highly inflected languages, it is uncommon to see all the forms of a word in the parallel corpora used during training. This problem is amplified for verbs in highly inflected languages where the correct form of the word depends on factors like gender, number and tense aspect. We propose a solution to augment the phrase table with all possible forms of a verb for improving the overall accuracy of the MT system. Our system makes use of simple stemmers and easily available monolingual data to generate new phrase table entries that cover the different variations seen for a verb. We report significant gains in BLEU for English to Hindi translation.

1 Introduction

Data driven approaches have become widely popular as they use little or no language specific knowledge. The main drawback of these approaches is the need for large amounts of data. (Koehn et al., 2003) have shown that the quality of the translations produced by data driven approaches mainly depends on the amount of parallel data available for the language-pair under consideration. Creation of a large bilingual corpus is expensive and time consuming if high quality manual translations are required. Hence, building MT systems for language-pairs with limited amounts of data is a big challenge.

Approaches have been suggested in the past to mine the world-wide-web to automatically obtain large amounts of parallel data. For example, news articles in two different languages describing the same event can be sentence-aligned to obtain a parallel corpus. Although this approach has shown improvements, this cannot be extended to languages that have little or no data on the world wide web.

The situation gets worse for languages that are rich in morphology. Clearly large amounts of parallel data are required to observe all variations of a word. Popovic and Ney (2004) applied transformations to verbs to reduce the number of out-of-vocabulary words and showed improvements in translation quality when morphemes were considered.

Yang and Kirchhoff (2006) used a back off model in a Phrase-based SMT system which translated word forms in the source language by hierarchical morphological abstractions. Unknown words in the test data were stemmed and phrase-table entries were modified such that words sharing the same root were replaced by their stems. Freeman et al. (2006) and Habash (2008) find in-vocabulary words for OOV words that could be morphological variants of the OOV words. Phrases in the phrase table containing these invocabulary words are then replaced by OOV words to create new entries. Vilar et al. (2007) used a letter-based MT system that treated the source and target sentences as a string of letters for translating unknown words.

All the above approaches handled OOV issues that arise when the source language is morphologically rich. Generation of the target sentence when the target language is morphologically rich from a source language that is not rich in morphology is non-trivial as the source language does not contain all the information for inflecting the target words. Minkov et. al (2007) predicted inflected forms of a sequence of word stems on languages that are morphologically rich using syntactic and rich morphological sources. This inflection generation model was then applied in MT by (Toutanova et al., 2008) while translating English into morphologically complex languages and showed improve-

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 111–119,
Chiang Mai, Thailand, November 8 – 13, 2011.

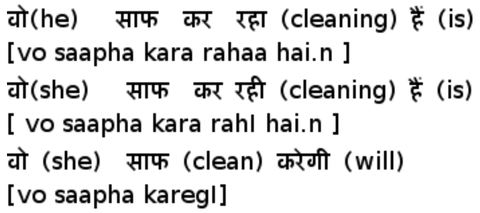
ment in translation quality. Their methods require a syntactic analyzer and a very rich morphological analyzer which may not be available for many rare or low-density languages. Also, their feature set includes bilingual features that require expensive and difficult to get bilingual corpora. We rely more on monolingual data and a small amount of parallel data. In cases of multi word compound words (explained in section 1.1) , since inflections on the light verb might change with change in the root verb compounding with it, we need to predict these verbs together and not as separate words.

In this paper, we consider Indian languages which are considered as low density languages as they do not have rich knowledge sources such as parsers or complex morphological analyzers. These languages also suffer from data sparsity and hence form ideal languages for the analysis of our proposed method. We also consider only various forms of verbs and do not consider other words such as noun phrases and adjectives affected by inflections.

1.1 Background on Indian Languages

India has fifteen official languages which originated from the Indo-Iranian branch of the Indo-European language family, the non-Indo-European Dravidian family, Austro-Asiatic, Tai-Kadai and the Sino-Tibetan language families (Microsoft Encarta Online Encyclopedia, 1997). The languages that stem from the Dravidian family, are - Tamil, Kannada, Malayalam and Telugu, spoken in the South Indian states. Languages in North India, such as Hindi, Urdu, Punjabi, Gujarati, Bengali, Marathi, Kashmir, Sindhi, Konkani, Rajasthani, Assamese and Oriya, stem from Sanskrit and Pali.

Indian languages are verb final i.e., verbs are placed at the end of the sentences. Verbs in these languages are inflected to contain information about gender (masculine and feminine), tense, aspect and number of the subject (singular or plural). A few examples showing inflections on the verbs in Hindi are shown below:



वो(he) साफ कर रहा (cleaning) हैं (is)
[vo saapha kara rahaa hai.n]
वो(she) साफ कर रही (cleaning) हैं (is)
[vo saapha kara rahI hai.n]
वो (she) साफ (clean) करेगी (will)
[vo saapha karegI]

These languages also contain compound verbs (multi-word compound representing a single verb). They contain a light verb which receives inflections and another component that can be a noun or a verb responsible for conveying the meaning. For example, in Hindi, most commonly used light verbs are "karna" (to make), "lena" (to take), "hona" (to happen) and "dena" (to give).

2 Motivation

When translating from a morphologically poor language such as English to any of the Indian languages, finding the right translation along with the inflections on the verbs becomes difficult, especially when the amount of bilingual data available is scarce. We try to use the pattern behavior of verbs to tackle this problem. Table 1 gives an example of hindi verbs classified according to their light verbs. Hindi side is transliterated for the sake of clarity. It shows how the verb phrase of one compound verb (clean) can generate verb phrase for words in the same group (help and forgive) just by replacing the corresponding source and target root words. The suffixes (shown in bold) are separated from the word to show how the actual process takes place. This paper tries to automatically group the different kinds of verbs occurring in the language based on their light verbs and generates the variation for all the verbs in one group by looking at the variations of any one member.

Karna	Lena	Dena
saaph (clean) maaph (forgive) madad (help)	sokh (absorb) bhaag (participate) goad (adopt)	saaza (punish) jawab (answer) anumati (allow)

he will be *clean* **ing** -> vo *saaph* kar **egaa**
he will be *forgive* **ing** -> vo *maaph* kar **egaa**
he will be *help* **ing** -> vo *madad* kar **egaa**

Table 1: Example of verbs belonging to different groups based on their light verb

The novel concept of this paper is generation of verb phrases on the source side and their translations using a) source and target monolingual data, b) simple morphological segmentation on source and target side and c) Small amount of manual translations or d) Word alignments of parallel corpora. We have described two methods of generating these verb phrase translations in the following sections.

3 Manual Generation

The idea here is to get the manual translations of source and target verb pairs which could capture the entire range of variations seen in the source and the target verb phrases. These translations can then be used to generate the variations for rest of the verb pairs. The entire flow of the method is shown in Figure 1.

3.1 Verb Phrase Chunking

Given a language pair (e,f), we extract all verb phrases that occur in the source monolingual data using a verb phrase chunker. Part of speech (POS) tags can be used to extract verb phrases for languages having a good POS tagger. In our experiments for English-Hindi language pair, POS tags were used for English verb phrase chunking. Modals in English were included as a part of the verb phrase since their counterparts in Hindi appear as verbs. For Hindi, the verb phrase chunker was trained on a small set of 6000 sentences, where the reference markings were obtained by projecting the verb phrases from English. The 6000 sentences were hand aligned with the corresponding English sentences, hence helping with the accuracy of the projected verb phrases. On this data, we built a CRF based chunker (Lafferty et al., 2001) using word and POS tag features.

3.2 Verb Classing

Using a segmenter, the root verb is separated from its inflected suffix for all the extracted verb phrases. These extracted verb phrases are then clustered based on the root verb so that all the variations of a root verb '<verb>' are grouped together into one cluster. As an example, a part of the verb cluster for 'play' is shown below. Note that all possible variations of each verb (both source and target side) are under one cluster.

play
was play +ing
should have play +ed
ought be played
would have been play +ed
is play +ed
cannot be play +ed
is being play +ed

The different variations within each verb cluster are normalized by replacing the root verb by a normalization tag '<verb>' so that similar root verb clusters now contain exactly the same variations and can be aggregated together easily. These clusters are put in N different 'verb classes' so that all verbs occurring with the same variations are under a single class and those with different variations are put in separate classes. For instance, since 'play' and 'help' have the same variations, they would belong to the same class. Choosing any one member of the class will cover all the variations of that class and by choosing one member from each of the N classes, all possible variations of all verbs in a given language are covered.

Class	Verb Class	**No of words**
AH	No Auxiliary	854
BH	'karna' as Auxiliary	1772
CH	'dena' as Auxiliary	212
DH	'lena' as Auxiliary	90
EH	'hona' as Auxiliary	242
	Rest	309

Table 2: Count of verbs belonging to different classes in Hindi

For English, we used morphA, (Minnen et al., 2000) an open source stemming package, to get the root form of the head verb in the extracted English verb phrases. It was observed that all the root verb clusters had the same variations of verb phrases and hence all belonged to the same class. Thus, from the source (English) side, only one verb could be picked up to cover all the variations, which could be then replicated for all the others verbs. We call that class 'AE'.

In Hindi, as explained in section 1.1, many verbs occur as compound verbs where a noun followed by a light verbs is considered as a verb and hence we also included these for our clustering and classification. The extracted verb phrases were segmented using a stemmer similar to one in (Ramanathan et al., 2003). After clustering the verb phrases based on their root verbs into groups and classing them based on the different variations, the main classes depended on the a) whether the verb has a light verb or not and b) the type of light verb attached. Table 2 shows the different classes found along with the number of verbs within each class.

There were more classes with a different auxiliary verb but we neglected them since the frequency of verbs in those classes was insignificant. 'lena' and 'dena' verb forms take the same variations (differing only in one character), we could easily generate one from another. Overall, only 3 classes were used for manual translation on the Hindi side(AH,BH and CH). Note that we allow a verb to belong to more than one class, suggesting that a word can be used in more than just one way depending on the context.

3.3 Root verb translation pairs

Given a parallel corpus, it is possible to extract source root verb to target root verb translation. Since a parallel corpus will contain the inflected form of a verb, it is necessary to stem them to their root form before calculating the word translation probabilities. Hence, given a parallel corpus, sentences are machine aligned by a maxent model as described in (Ittycheriah and Roukos, 2005) and then the verbs on both the source and target side are stemmed to the respective root forms using a suitable stemmer. Given the list of possible source verbs and target words from the previous clustering step, the forward and reverse translation probabilities for these verbs is calculated from the alignments using by relative frequency:

$$P(f_i/e_j) = count(f_i, e_j)/\sum_f count(f, e_j) \quad (1)$$

$$P(e_i/f_j) = count(e_i, f_j)/\sum_e count(e, f_j) \quad (2)$$

Using these forward and reverse probability, a mapping file that maps the source root verb to the corresponding target root verb(s) is created by empirically combining the two probabilities.

$$P_{tot}(e_i/f_j) = 0.5{*}P(e_i/f_j){+}0.5{*}P(f_j/e_i) \quad (3)$$

We allow one source verb mapping to multiple target verb, since the meaning can change due to context in the test sentence. TopM translations of source word are selected for translations, provided $\mathrm{P}_{tot} > \mathrm{P}_{thresh}$. We empirically found the P_{thresh} = 0.2 and M=4 to work reasonably well. Two or more worded root verb, such as phrasal verbs ‘take off’, ‘figure out’, were not considered while creating the mapping since the meaning is often different that the individual words and the generation of verb phrases from these root verbs is more tricky. Such constructions, where one word verb may translate to multiple words, occurred for only 3% of the verbs in the test data and hence could be ignored without any significant loss in improvement.

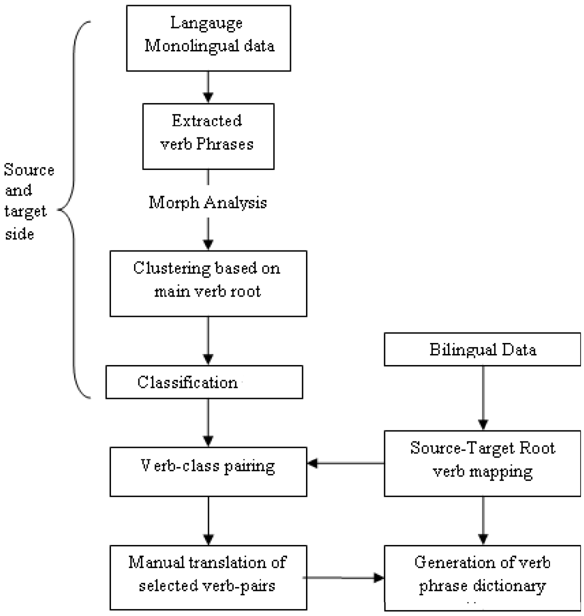

Langauge Monolingual data
Extracted verb Phrases
Morph Analysis
Clustering based on main verb root
Classification
Source and target side
Bilingual Data
Verb-class pairing
Source-Target Root verb mapping
Manual translation of selected verb-pairs
Generation of verb phrase dictionary

Figure 1: Steps involved in manual generation of verb phrases

3.4 Generation of Verb Phrase Dictionary

Given the root verb mapping and the classes to which these source and target root verb belong to, we create a ‘source class’ to ‘target class’ mapping, or a ‘verb-pair class’, by replacing the root verbs with their corresponding verb classes. This causes each of the verb pair to fall under some particular verb-pair class. If there are n classes in the source side and m on the target side, the maximum number of verb-pair classes are

N = m*n

By picking any one root verb pair from each of these verb-pair classes, we can cover all the possible variations of verb phrase translation pairs. These pairs can then be given for human translation, by creating all possible variations of either the source side or target side and asking humans to translate to the other.

Templates for each of the N verb-pair class are created from the manually translated data by segmenting the verb phrase pairs on both sides and replacing the root verb by the ‘<verb>’ tag. An example of such translation pair for English-Hindi is shown

was <verb> +ing == <verb> raha tha [Class AH-AE]
was <verb> +ing == <verb> kar raha tha [Class BH-AE]

Picking root verb pairs from each verb-pair class and replacing the <verb< tag with corresponding verbs, these templates are used to create new verb phrases which may not be present in the parallel data to a large extent. A reverse morphological tool or joiner is used to recombine the segmented verb phrases and create a verb phrase dictionary.

In our paper, English had one class and Hindi had 3. Thus, only 3 Hindi-English verb pairs needed to be translated, one from each of the verb-pair classes AH-AE, BH-AE, and CH-AE. We created different variations of the English verbs, since it had only one class and could be easily generated using manually built rules. Grammar rules contain number (singular/plural), tense and aspect agreement between different auxiliary forms (for example: was, is, were, can, might, could not, wouldn't, etc) and verbs (for example: answer, punishing, cleaned, etc.). A unique Hindi-English verb pair is picked from each of the verb-pair classes obtained earlier and their English verbs are used in generating the English verb forms. For example, "saaf" belonged to the "karna" cluster, so its English translation, "clean" is used for creating verb forms. Gender information is also added to the Verb forms which will be required for the Hindi counterparts. About ≈ 970 verb forms were generated for each of the 3 verbs. Examples of a few Verb forms are given below:

[he] will clean
[we] will clean
[I(he)] will clean
[he] may not have been cleaning
[she] could have been cleaning
[we] may have been cleaning
[I(she)] may have cleaned

'Not' is included as a part of the extracted verb phrases since it is the most common adverb that occurs within the verb phrases. Other adverbs such as 'now', 'also' have not been dealt with in this paper. These variations, along with the mapping of the English-Hindi root verb was given to the annotators for translation. The subject information within '[]' helps the annotators to decide on the number and gender inflections on the target (Hindi) side. These are removed before using in the machine translation system. The reverse morphology of the generated verb phrases is done using MorphG (Minnen et al., 2000) for English and a simple suffix joiner for hindi.

4 Automatic generation

Although manual translation is a clean and effective way of generating these verb phrases, a human is still required in the loop to complete the setup. Instead, both side monolingual data can be employed to extract all the variations for each root verb, parallel corpus can be used to get the source-verb to root-verb pairs, and finally a model can be learnt to align the source verb phrase to target verb phrase using the verb alignments from the hand alignments and machine alignments.

4.1 Verb pairs

Using the technique described in section 3.3, a source to target root verb mapping are obtained.

4.2 Clustering

Clustering of verb phrases on source and target side is done as explained in section 3.2 so that each cluster contains different variations of the same root verb form. The phrases within each cluster are segmented on both sides using the techniques described section 3.2 and are generalized by replacing the root verb in the segmented verb phrase by a '<verb>' tag as this will help while aligning the source verb phrase to its corresponding target verb phrase.

4.3 Verb Phrase Translations

In order to learn a verb phrase alignment model, we need good quality verb phrase alignments from the parallel corpora. We concentrate on hand aligned data and accurate machine alignments. Machine aligned verb phrases that occurred less than three times were treated as inaccurate. The source side verb phrases are extracted using the scheme similar to one in section 3.1, and by looking at the target words they align to, verb phrase alignments are obtained. The aligned verb phrases are segmented on both the target and the source side using the strategy described in section 3.2 and then normalized by replacing the head word for both the source side verb phrase and the target side verb phrase by a '<verb>' tag. The '<verb>' tagged verb phrases act as templates for verb alignments. Since all the root forms of verb will not occur in the extracted verb alignments, it's necessary to normalize them to be able to learn a general model. This way, if the translation of a particular source verb phrase variation is known, its generalized form can be used to get the translation of a different root verb for the same variation. This is similar to our claim in section 3.4 that translation of one root verb can generate translations for all other verbs belonging to the same class.

A simple word alignment model is used to learn the word translation probabilities. Please note that the suffixes of the verbs are also treated as words since they contain important information about tense, gender and number. We used GIZA++ model 4 to learn P(V_{si}/V_{tj}), which is the probability of the i^{th} source word/segment aligning to the j^{th} target word/segment.

4.4 Automatic Alignment and Generation

From the root-verb pairs obtained in section 4.1, each verb pair is picked and the best translation for a source verb phrase in source-side verb cluster is searched for in the target verb cluster. If a cluster for the source or target verb does not exist, that pair is ignored. Both the source and the target verb clusters contain the generalized verb phrase of the form '... aux_{-1} <verb> aux_1 aux_2.. '. First, a perfect match of a source phrase and target verb phrase is searched in the hand aligned and machine aligned verb phrase pairs. If found, that phrase pair is treated as a valid verb phrase pair. If no perfect match is found, word alignment probabilities obtained in previous section are use to get the source to target verb phrase alignments. Any verb phrase alignment pair with score lower than a threshold score of 0.5 is ignored.

After obtaining all the valid verb phrase pairs, the tag <verb> is replaced by their corresponding root verbs and as in section 3.4 and the suffixes are joined to the root verb to get the automatically generated verb phrase dictionary which can be used in the MT system.

5 Experiments

In this section, we report our experimental results on English - Hindi language pair. We first report on the coverage ratio, which gives an estimate of number of exact verb phrases covered by the baseline system and our method. In addition, we also report on English to Hindi Machine translation results for phrase based systems.

5.1 Discovery of new data

The data used for clustering and classification on source and target side, the parallel corpora and the test set details are shown in table 3.

Data	No of Sentences
English Monolingual	6 million
Hindi Monolingual	1.4 million
Test set 1	4000
Test set 2	715
Training Data	280k

Table 3: Data used for experiments

The Hindi monolingual data was used to collect 4320 Hindi verb clusters belonging to 3 different classes (section 3.2) and the English monolingual yielded 4872 clusters. Many of these clusters were false positives due to the bad quality of verb phrase chunker but were eliminated in the subsequent steps. The parallel data was aligned using a maxent model (Ittycheriah and Roukos, 2005) and gave us 2944 verb-pairs.

For manual generation method, 3 verb pairs were given to annotators for translation, with about 972 different English verb forms in each. The generated dictionary from the manual translations had 2.7 million verb phrases. The automatic method aligned the corresponding clusters of the 2944 verb-pairs and produced about 300k new verb phrases. The considerably lesser size of the automatically created verb phrase dictionary compared to the manual dictionary can be attributed to the fact that the manual dictionary contains variations that are not seen in our monolingual data.

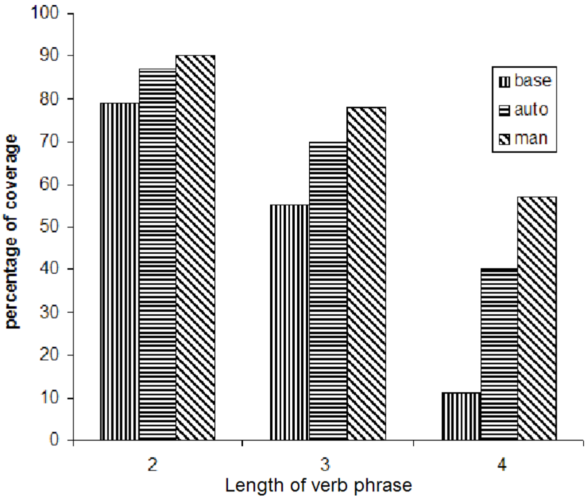


Figure 2: Coverage percentage for different settings

We claim that the manually and automatically created verb phrase dictionaries add new data to the system and have a higher chance of finding a matching source verb phrase in a given corpus than our phrase based system. We verify this claim by extracting verb phrases from two test sets and searching for them in:

1. Baseline Phrase table
2. Base+Manual generated Dictionary
3. Base+Auto generated Dictionary

Figure 2 shows the ratio of the number of verbs phrases found in the three cases to the total number of verbs searched. We call this as coverage ratio. The verb phrases are divided based on their lengths. The plot clearly shows that the coverage increases considerably by the addition of these generated verb phrases which may or may not be seen in the training data, especially as the length of the verb phrase increases. Verb phrases of length 1 are not shown since the coverage was almost same for the 3 settings.

5.2 Machine translation Results

Table 3 shows the training data and the two test sets used for evaluation. The bilingual parallel data is split into training data and Test set 1. Test set 2 is a generic test set. All the data (training and test) used is predominately contains news. We report are results on BLEU (Papineni et al., 2002).

The verb phrase pairs were generated as explained in manual generation section and then added to the baseline system as a part of corpus(Base+manVPcorp). To emphasis on the improvement from generated verb phrases, an experiment where only the human translated verb phrases are added to the baseline corpus was also conducted(Base+humanVPcorp).

Table 4 shows the results on Moses - a state of the art phrase based system, and on a phrase based system (PBMT) similar to (Tillman et al., 2006) on test sets 1 and 2. Both systems were trained on 280k parallel sentences. On the in-domain data, we had an improvement of 4.8 BLEU points for Moses and 0.9 for PBMT. On the more generic test set, Moses gave a BLEU score improvement of 1.1 whereas the PBMT performance was comparable. One reason for this difference in the BLEU score jumps is the better alignments in the PBMT system, aligned by a maxent model as described in (Ittycheriah and Roukos, 2005). The PBMT system thus has a higher chances of having a good verb phrase in the baseline system than Moses and hence on adding generated verb phrases, we would see a lesser gain. Since the PBMT system had comparable results on the in-domain data with Moses and performed better on the out of domain (more generic) test set, the remaining experiments have been conducted with the PBMT system.

	Moses		**PBMT**	
	Set 1	Set2	Set 1	Set2
Baseline	13.5	08.6	13.4	**16.1**
Base+humanVPcorp	14.1	08.6	14.0	16.1
Base+manVPcorp	**18.3**	**9.7**	**14.3**	16.0

Table 4: BLEU score on test set 1 and 2 for different settings on moses and PBMT

Adding the generated phrases as a parallel corpus can alter the translation probabilities of individual words and sub-phrases. This is one of the reasons for no improvement in the bleu score of the PBMT system when the generated verb phrases are added as corpus. A better method would be to add the verb phrases directly to the phrase table. We added the manual dictionary to the PBMT system and the results are tabulated in table 5.

	Set 1	**Set2**
Baseline	13.4	16.1
Base+humanVP-PT	14.0	16.1
base+manVP-PT	**14.9**	**16.5**
base+autoVP-PT	14.8	16.3

Table 5: BLEU score for PBMT system after adding verb phrases directly to Phrase Table (PT)

Adding the verb phrases directly to the system keeps the rest of the phrases and their scores intact. Only phrases with matching source side phrase need to be re-normalized to adjust the translation probabilities. This would mean that the probability of only the verb phrases we add to the baseline phrase table would be affected while the rest of the translation model would be the same. Table 5 shows that addition of manually generated data to the phrase table(Base+manVP-PT), gives a good improvement of 1.5 points on the in-domain data and 0.4 on the out of domain data. A significant improvement of 1.4 BLEU points is seen even when the automatically generated verb phrases are added (Base+autoVP-PT), which was not seen when these were added as a corpus to the system.

Figure 3 shows the variation of BLEU score with change in the corpus size. We should expect that the gain be higher in the case of low corpus size. However, note that the verb-pair list used to generate the verb phrases also changes with the change in corpus size, since decreasing the corpus size would decrease the quality of the overall alignments and the number of verbs seen. Thus, while the verb-pair list using 160k sentences had a total of 2499 verb pairs, the 20k corpus produced only 1347 verb pairs. So, for a smaller corpus size, the number of new verb phrases added to the table would also be lesser. This explains the rather constant gain in BLEU score throughout the graph.

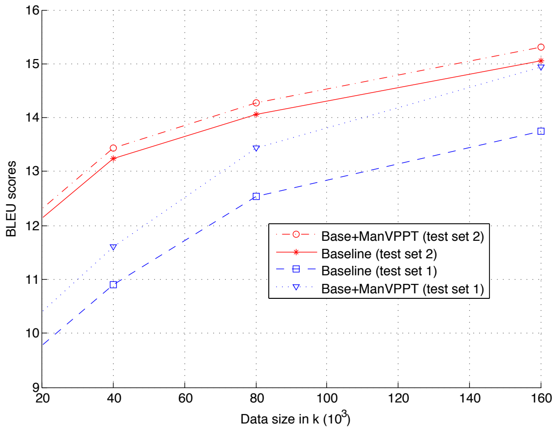


Figure 3: Change in BLEU for different corpus sizes

6 Conclusion and Future Work

We showed an improvement of up to 1.5 bleu points on the in domain data set and an improvement of 0.4 bleu on the generic test set. However, there are still some errors in respect to the morphology of the verb phrases, which the Language Model is unable to tackle. These are primarily the long range dependencies which includes determining the gender and number of the subject or object to get the appropriate inflection. Having a dynamic feature based system, which does not require rich morphological resources, and predicts the suffixes and inflections would be able to solve this problem. Also, when a verb has more than one meaning, the contextual information is not captured efficiently in the current method and often produces a more literal translation than the the reference.

Apart from adding the verb phrases to the phrase table, filtering of poor verb phrase pairs from the original phrase table is another approach to consider. The two methods together can give a higher boost to the translation than just one of them. A more language independent method of extraction of verb phrases also needs to be constructed, which does not require building language dependent stemmers and verb phrase chunkers.

References

Einat Minkov, Kristina Toutanova and Hisami Suzuki. 2007. *Generating Complex Morphology for Machine Translation*, in Proc. 45th Annual Meeting of the Association for Computational Linguistics, 2007, pp. 128-135.

Kristina Toutanova, Hisami Suzuki and Achim Ruopp 2008. *Applying Morphology Generation Models to Machine Translation*, in Proc. 46th Annual Meeting of the Association for Computational Linguistics, 2008.

D. Vilar, J. Peter, H. Ney, and L. F. Informatik 2007. *Can we translate letters?*, In Proceedings of Association Computational Linguistics Workshop on SMT, pages 33-39, 2007.

A. T. Freeman, S. L. Condon, and C. M. Ackerman 2006. *Cross linguistic name matching in english and arabic: a "one to many mapping" extension of the levenshtein edit distance algorithm.* , In Proceedings of the main conference on Human Language Technology, Conference of the North American Chapter of the Association of Computational Linguistics.

N Habash 2008. *Four techniques for online handling of out-of-vocabulary words in arabic-english statistical machine translation* , In Proceedings of Association for Computational Linguistics-08.

M. Popovic and H. Ney 2004. *Towards the use of word stems and suffixes for statistical machine translation*, In Proceedings of The International Conference on Language Resources and Evaluation.

M. Yang and K. Kirchhoff 2006 *Phrase-based backoff models for machine translation of highly inflected languages*, In Proceedings of the European Chapter of the ACL, pages 41-48, 2006.

Philipp Koehn, Franz Josef Och, Daniel Marcu. 2003. *Statistical Phrase-Based Translation*, In Proceedings of HLT-NAACL 2003.

Eleftherios Avramidis, Philipp Koehn 2008. *Enriching Morphologically Poor Languagesfor Statistical Machine Translation* , In Proceedings of ACL-08, HLT.

Ananthakrishnan Ramanathan and Durgesh Rao 2003. *A Lightweight Stemmer for Hindi 2003*, Workshop on Computational Linguistics for South-Asian Languages, EACL.

Ananthakrishnan Ramanathan, Pushpak Bhattacharyya, Jayprasad Hegde, Ritesh M. Shah., Sasikumar M 2007 *Simple Syntactic and Morphological Processing Can Help English-Hindi Statistical Machine Translation*, In Proceedings of International Joint Conference on Natural Language Processing,2007.

Christoph Tillman 2006 *Efficient Dynamic Programming Search Algorithms for Phrase-based SMT*, In Proceedings of the Workshop CHPSLP at HLT'06.

Minnen, G., J. Carroll and D. Pearce 2000 *Robust, applied morphological generation*, In Proceedings of the 1st International Natural Language Generation Conference, Mitzpe Ramon, Israel pp 201-208.

Abraham Ittycheriah and Salim Roukos 2005. *A maximum entropy word aligner for arabic-english machine translation*, In Proceedings of HLT/EMNLP, HLT-05, pages 89-96, Stroudsburg, PA, USA. Association for Computational Linguistics.

John Lafferty, Andrew McCallum, and Fernando Pereira 2002. *Conditional random fields: Probabilistic models for segmenting and labeling sequence data* , in Proc. International Conference on Machine Learning (ICML), 2001.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. 2002. *BLEU: a method for automatic evaluation of machine translation* , in Proc. ACL-2002: 40th Annual meeting of the Association for Computational Linguistics pp. 311 - 318.