SMT from Agglutinative Languages: Use of Suffix Separation and Word Splitting

Prakash B. Pimpale
KBCS, CDAC Mumbai
prakash@cdac.in

Raj Nath Patel
KBCS, CDAC Mumbai
rajnathp@cdac.in

Sasikumar M.
KBCS, CDAC Mumbai
sasi@cdac.in

Abstract

Marathi and Hindi both being Indo-Aryan family members and using Devanagari script are similar to a great extent. Both follow SOV sentence structure and are equally liberal in word order. The translation for this language pair appears to be easy. But experiments show this to be a significantly difficult task, primarily due to the fact that Marathi is morphologically richer compared to Hindi. We propose a Marathi to Hindi Statistical Machine Translation (SMT) system which makes use of compound word splitting to tackle the morphological richness of Marathi.

1 Introduction

Marathi is widely spoken in and around Maharashtra, India and also in other parts of the world. Hindi is widely spoken in Northern India and is understood in most parts of the nation. Hindi also has significant number of speakers across the world in countries where Indians have migrated. Marathi speaking areas host many important economic and social activity centers, where many times there is need for translation of content from Marathi to Hindi.

Marathi and Hindi both belong to the Indo-Aryan family of languages and have the same flexibility towards word order, canonically following the SOV structure. As both are written in Devanagari script and have many words which are either same or can be traced to same origin, they resemble each other to a great extent. This resemblance may make us to believe that Statistical Machine Translation will be an easier affair on this pair. But upon experiments it is observed that the morphological richness of Marathi makes it as difficult as any other Indian language to Indian language Machine Translation.

Marathi is agglutinative in nature which makes Marathi to Hindi SMT even more difficult. It is known that SMT produces more unknown words resulting in bad translation quality, if morphological divergence between source and target languages is high. Koehn & Knight (2003), Popovic & Ney (2004) and Popovic et al. (2006) have demonstrated ways to handle this issue with morphological segmentation of words before training the SMT system.

We demonstrate a better performing Marathi to Hindi SMT system which makes use of morphological segmentation on the source side prior to training. The proposed system shows significant improvement in translation quality compared to the baseline. We also present comparative study using BLEU (Papineni et al. 2002), NIST (Doddington, 2002), Position-independent Word Error Rate (Tillmann et al., 1997), Word Error Rate (Nießen et al., 2000), manual evaluations and 10-fold cross validation.

The rest of the paper is organized as follows. In Section 2, we discuss the similarities and dissimilarities in the language pair under study. In Section 3, we describe the experimental set up and splitting algorithm. Section 4 discusses experiments and results for splitting and constrained splitting. Analysis and discussion is done in section 5 with manual evaluation, 10-fold cross validation, error analysis and comparative study with similar work, followed by conclusion and future work in section 6.

2 Similarity Analysis of Marathi and Hindi

Marathi and Hindi both belong to Indo-Aryan family of languages and Marathi is Southern-most in this category. Being situated in such a geographical vicinity of India Marathi seems considerably influenced by Dravidian languages (Junghare, 2009). It makes frequent use of word compounding or post modifications to create meaningful words using prefixes and suffixes. Number of such derived words in Marathi is very high and this distinguishes Marathi from others in Indo-Aryan language family. Dabre et al.

(2012) and Bhosale et al. (2011) have theoretically discussed the morphological richness of Marathi and compared it with Hindi.

We analyzed a parallel Marathi and Hindi translation corpus of size 48000 sentences for following parameters:

- Average Sentence Length: Considered number of words in a sentence as the sentence length. This parameter is captured to compare number of words a language needs to represent a concept, assuming a sentence is written to represent a concept.
- Word Count: Total number of words in the corpus. This is captured to affirm that a morphologically poorer language needs more words as compared to the richer.
- Unique Word Count: Number of distinct words in the corpus. This is computed to compare morphological richness of the languages.
- Average Word Frequency: Word frequency is number of times the word is repeated in corpus. This will help to demonstrate that word frequency is higher in morphologically poorer language.
- Average Word Length: Number of characters in a word is word length. This is measured to analyze the significant presence of compound words in Marathi.

The corpus analysis in Table 1 shows that a Hindi sentence needs on an average 17 words to represent a concept whereas Marathi sentence needs just 12 words to represent the same concept. The total number of words in Hindi corpus is 834417 and Marathi has just 602500 which affirm the fact that Marathi represents varied concepts with lesser number of words as compared to Hindi. We can also see that unique word count for Marathi is more than Hindi by 44474, which demonstrates that Marathi has larger vocabulary of surface forms to describe different meanings, and the same in case of Hindi is done by using different word combinations. As Hindi has less unique words it needs to repeat many of them for representing certain meanings and this is evident from the higher average word frequency of 19. Marathi has comparatively less word frequency as it doesn't need to do the same. The average word length for Marathi is higher and it shows that significant number of Marathi words carry more information than their Hindi counterparts. The length difference also demonstrates that the compound words are significantly high in Marathi and thus statistically affirms the morphological richness of Marathi compared to Hindi.

In Marathi there are words like 'हरिद्वारमध्येही' (*haridwarmadhyehi* – *a*lso in Haridwar) and 'पोहोचण्याकरिता' (*pohachanyakarita* – to reach) which when translated to Hindi will become 'हरिद्वार में भी' (h*aridwar men bhi*) and 'पहुंचने के लिए' (*pahunchane ke liye*) respectively. Here the word 'हरिद्वारमध्येही' (*haridwarmadhyehi*) is formed by compounding a proper noun 'हरिद्वार' (h*aridwar*), preposition 'मध्ये' (*madhye* - in) and an adverbial 'ही' (*hi* - also) and 'पोहोचण्याकरिता' *(pohachanyakarita)* is formed by compounding 'पोहोचण्या' (*pohachanya* - derived verb form of 'reach') and 'करिता' (*karita* - 'TO' infinitive equivalent in Marathi). Marathi follows different rules for derivation of such words by stacking together different surface forms and suffixes. In the process (called as *Sandhi*), it may modify the form of surface word.

As an example we can see word 'उपाहारगृहाप्रमाणे' *(upahargruhapramane)* is formed by combining 'उपाहारगृह' (*upahargruh* - restaurant) and 'प्रमाणे' (*pramane* – as per)*;* but while combining these two words, ' ा' (*aa*) letter is attached to 'उपाहारगृह' *(upahargruh)* as suffix to derive a new base form 'उपाहारगृहा'

	Average Sentence Length	**Word Count**	**Unique Word Count**	**Average Word Frequency**	**Average Word Length**
Hindi	17	834417	43342	19	6
Marathi	12	602500	87816	6	8

Table 1. Analysis of Marathi-Hindi Parallel Corpus (48000 Sentences)

Marathi Word	Hindi Translation	English Translation
उपाहारगृहाप्रमाणे *(upahargruhapramane)*	भोजनालय के अनुसार *(bhojanalay ke anusar)*	As per the restaurant
रेल्वेमार्गावर *(relwemargavar)*	रेल मार्ग पर *(rel marg par)*	On the railway route
परिवर्तनाशिवाय *(pariwartanashivay)*	परिवर्तन के सिवाय *(pariwartan ke siway)*	Without changes
त्यावेळेपर्यंत *(tyaveleparyant)*	उस समय तक *(us samay tak)*	By that time
घरापासून *(gharapasun)*	घर से *(ghar se)*	From home

Table 2. Marathi to Hindi Translation Examples

Common Words in Hindi and Marathi - CW	CW as % of Hindi Unique words	CW as % of Marathi Unique words
16693	38.51	19.00

Table 3. Marathi-Hindi Parallel Corpus Similarity Analysis

	Number of Bi-lingual Sentences	Number of Hindi Words	Number of Marathi Words
Training(TM)	49000	854995	644878
Training(LM)	72394	1475217	-
Testing	1000	17660	13372

Table 4. Corpus Distribution, TM-Translation Model, LM- Language Model

```
BEGIN
   INITIALISE suffixSet
   INITIALISE splits = {candidateWord, “NULL”}
   FOR suffix IN suffixSet:
      IF candidateWord ENDSWITH suffix AND candidateWord.LENGTH > suffix.LENGTH
      splits[0] =  candidateWord.SUBSTRING(0, candidateWord.LASTINDEXOF(suffix))
      splits[1] = suffix
      RETURN splits
END
```

Figure 1. Splitting Algorithm

(upahargruha) and then word 'प्रमाणे' (*pramane*) is suffixed. More examples demonstrating this phenomenon have been provided in Table 2. We can observe that a single word in Marathi is translated to multiple words in Hindi and English and that's due to the morphological richness of Marathi compared to Hindi and English.

Another analysis presented in Table 3 shows that Marathi and Hindi have around 16693 words in common which are 38.51% of Hindi and 19.00% of Marathi vocabulary extracted from the corpus. Some of these words are common nouns like 'सचिव' (*sachiv*) and 'चित्र' (*chitra*), proper nouns like 'आकाश' (a*kash*), 'नागापट्टिनम' (n*agapattinam*) and a few words from other languages transliterated in Devanagari script like 'इन्सुलिन' (insulin) and 'टेक्नोलॉजी' (technology). Many of these common words have their origin in Sanskrit and are used as it is or on derivation. We also need to notice that the foreign language words transliterated into Devanagari are part of this common words set as both the languages use Devanagari for representation.

3 Experimental Setup

In the following subsections we describe training corpus and SMT system setup for the experiments.

3.1 Corpus for SMT Training and Testing

A prime need for any SMT system is good quality bi-lingual corpus. We have used manually translated bi-lingual corpus of size 49000 sentences for training the translation model. The 49000 bi-lingual corpus of Health and Tourism domains contained 854995 Hindi words and 644878 Marathi words. Language model training was done using monolingual Hindi corpus of size 72394 sentences. A set of 1000 unseen sentences has been used for testing the systems. The test set contained 500 sentences from Health and Tourism each. Table 4 summarizes the Training and Testing data.

3.2 Splitting Marathi Words

To tackle the described morphological complexity of Marathi for the purpose of better SMT system we have devised an algorithm to split inflected and compound Marathi words. The splitting algorithm uses a list of suffix and commonly compounded words as suffixes, combinedly referred as suffix list hereafter. The list is created from the available bi-lingual and monolingual corpus.

3.2.1 Creating Suffix List

To develop the Marathi Splitter we trained an alignment (GIZA++; Och and Ney, 2003) model to get the Marathi-Hindi phrase alignments. Upon training we extracted Marathi words which align to multiple Hindi words from the alignment table. The extracted Marathi word set was then manually analyzed to develop a list of valid compound words. From the list of valid compound words, we further extracted high frequency suffixes. On manual analysis of these suffixes a valid list of suffixes for splitting (list 1) was developed. We also analyzed the Marathi corpus and extracted words with length more than 10 (as the average word length for Marathi is 8). These extracted words were then manually analyzed to get a comprehensive list of compound suffixes (list 2). The final set of 129 suffixes was a combination of list 1 and list 2.

3.2.2 Splitter Algorithm

The algorithm splits a given Marathi word if it contains a suffix from the list created. Figure 1 shows pseudo-code for Marathi Splitter. The algorithm will split Marathi word 'उपाहारगृहासारखी' (*upahargruhasarkhi* – like restaurant) into 'उपाहारगृह' (*upahargruha* – restaurant) and 'ासारखी' (*aasarkhi*) which are valid and invalid dictionary words respectively. In case of 'ासारखी' *(aasarkhi)*, 'ा' *(aa)* is a *Sandhi* marker and 'सारखी' (*sarkhi*) means 'like' in English. Word 'प्रसारमाध्यमकेंद्र' (*prasarmadhyamkendra* – media center) on splitting will give 'प्रसारमाध्यम' (*prasarmadhyam* - media) and 'केंद्र' (*kendra* – center) which are valid dictionary words. Though most of the splits give at least one valid dictionary word, there are cases where it fails to do so. Like in case of 'घेण्याचा' (*ghenyacha* – to take), the splits will be 'घेण्य' (*ghenya*) and ' ाचा' (*aacha*), where both are invalid dictionary words.

3.3 SMT System Setup

The baseline system was setup by using the phrase-based model (Och and Ney, 2003; Brown et al., 1990; Marcu and Wong, 2002; Koehn et al., 2003) and Koehn et al. (2007) was used for factored model. The language model (5-gram) was trained using KenLM (Heafield, 2011) toolkit with modified Kneser-Ney smoothing (Chen and Goodman, 1998). For factored SMT training source and target side stem has been used as alignment factor. Stemming has been done using Ramanathan and Rao (2003) lighweight stemmer for Hindi. The stemmer for Marathi has been developed by modifying Ramanathan and Rao (2003).

3.4 Evaluation Metrics

The different experimental systems have been evaluated using, BLEU (Papineni et al., 2002), NIST (Doddington, 2002), position-independent word error rate (Tillmann et al., 1997), word error rate (Nießen et al., 2000) and manual evaluations. For a MT system to be better, higher BLEU and NIST scores with lower position-independent word error rate (PER) and word error rate (WER) are desired.

4 Experiments and Results

In the following subsections we discuss different SMT systems and their performance. We also study the impact of splitting on output of SMT systems. Further we discuss methodologies to improve splitting and hence the translation quality.

4.1 Impact of Splitting

For training the translation model we used 49K bi-lingual corpus and language model was developed using 72.394K Hindi sentences. We used splitting discussed in section 3.2.2, as a pre-processing step for training phrase-based and factored SMT systems, MH3 and MH4 respectively. The systems are described in Table 5.

Results for the systems described in Table 5 are detailed in Table 6. Impact of splitting can be observed by comparing MH1 and MH3. We also notice that factored systems, MH2 and MH4 are performing better than phrase-based systems, MH1 and MH3 respectively. The significant improvements in all evaluation metrics demonstrate that splitting of Marathi words is helping to achieve better translation quality.

	MH1	**MH2**	**MH3**	**MH4**
BLEU	38.35	38.55	41.71	**42.01**
NIST	7.756	7.778	7.982	**8.023**
PER	42.08	42.15	39.58	**39.28**
WER	35.82	35.61	32.21	**31.91**

Table 6. Experiment Results (in %)

Upon analysis of the translations by MH4, we noticed that some of the words were getting wrongly translated. For example 'वरात' (*varat* – A marriage function) which should not have been split, was split into 'वर' (*var*) and ' ात' (*aat*) resulting into incorrect translation as 'पर हैं' (*par hai* – over/at). How to tackle such errors? Can we use length constraints to prohibit such words from splitting? Can POS (NNP) constrain help? These questions lead us to investigate further. We experimented various combinations of length and POS constraints which are described in following section.

4.2 Constrained Splitting

We use splitting discussed in section 3.2.2, as a pre-processing step for training various phrase-based and factored SMT systems. However, we apply constraints over word length and POS tag before splitting. The systems with different constraints are described in Table 7. For MH5, MH7, MH8 and MH9, words with character length at least five were considered for splitting. This particular length constraint was selected as it gave maximum BLEU score, on experimenting with different lengths ranging from 4 to 8. For MH5 and MH6 pre-processing was performed only once. To tackle words like 'आजारापासूनसुद्धा' (*aajaranpasunsuddha*), formed by compounding N multiple words, they need to be split N-1 times. We have tried to handle these cases in MH7, MH8 and MH9 by two level and multi-level splitting as detailed in Table 7. In MH7 a word was subjected to pre-processing twice, whereas in case of MH8 and MH9, the same was done as long as the word satisfies length criterion. Further, with the aim to prohibit splitting of named entities like 'परमेश्वर' (*parmeshwar*), 'पेशावर ' (*peshawar*) and 'खरात ' (*kharat*), we tried applying NNP POS tag constraint in MH6.

System	**Description**
MH1	Phrase-Based SMT System (Baseline)
MH2	Factored SMT System
MH3	Phrase-Based SMT System with Splitting (all words considered as candidates for splitting)
MH4	Factored MH3 (stem as alignment factor on source and target side)

Table 5. System Description

System	**SMT Model**	**Splitting Candidate Criteria for A Word**	**Splitting Level**
MH5	Phrase-Based	A word with character length >= 5	One
MH6	Phrase-Based	All words except proper nouns (NNP)	One
MH7	Phrase-Based	A word with character length >= 5	Two
MH8	Phrase-Based	A word with character length >= 5	Multi
MH9	Factored	A word with character length >= 5	Multi

Table 7. System Description

	MH3	MH4	MH5	MH6	MH7	MH8	MH9
BLEU	41.71	42.01	41.70	41.24	41.94	41.93	**42.06**
NIST	7.982	8.023	7.987	7.953	8.025	8.023	**8.029**
PER	39.58	39.28	39.53	39.83	39.25	39.20	39.26
WER	32.21	31.91	32.19	32.63	32.07	32.04	31.88

Table 8. Evaluation (in %)

Criteria	**% Accuracy**	**Grade Scale**
Syntactically well-formed / semantically high acceptance	80% and above	4 point grade scale
Syntactically well-formed / semantically low acceptance	60% - 79%	3 point grade scale
Syntactically well-formed / semantically unacceptable	40% - 59%	2 point grade scale
Syntactically ill-formed / semantically unacceptable	below 40%	1 point grade scale
No output / garbage output	-	0 point grade scale

Table 9. Grading Scheme

Table 8 details the results obtained on different evaluation metrics for the experimental systems. We can see that among all, the highest BLEU and NIST scores are achieved by MH9 which is factored SMT system and makes use of length constrained multi-level splitting. There is not much difference in BLEU for MH3 and MH5. But MH7 and MH8 show significant improvement in BLEU over MH3. BLEU for MH6 is slightly decreased, as many words like 'राजस्थानात' (*rajasthanat* – in Rajasthan) which are candidates for splitting are not getting split because of their NNP POS tag. Use of a Marathi NER may be experimented to tackle this issue in future. In next section, we have further analyzed and compared manual evaluation and 10-fold cross validation for some of these systems to better understand the performance difference.

5 Analysis and Discussion

We discuss here, manual evaluation, 10-fold cross validation and error analysis followed by comparative study with the existing work. MH1, MH3, MH5, MH8 and MH9 only have been considered for manual evaluation, as comparison of these systems is sufficient to understand the contribution of splitting to translation quality.

5.1 Manual Evaluation

Figure 2 shows manual evaluation of systems (MH1, MH3, MH5, MH8 and MH9) for 50 random sentences from the test set. For the evaluation, sentences were translated using systems under study and graded as per the scheme detailed in Table 9.

Figure 2 shows that among the systems compared, MH9 has highest number of sentences with accuracy more than 80%. We can also see that use of constraints on splitting in MH5 has helped reduce the number of sentences in grade 2 as compared to MH3. That shows, semantic acceptance of translations is increasing with the use of constrained splitting.

Table 11 describes with the help of an example, improvement in the quality of translation upon use of splitting. In the input sentence, words 'वलसाडच्या' (*valsadchya* – of Valsad) and 'किनाऱ्यावर' (*kinaryavar* – *o*n the bank) are candidates for splitting. These words are split into 'वलसाड' (*valsad*) + 'च्या' (*chya*) and 'किनाऱ्या' (*kinarya*) + 'वर' (*var*), respectively. We can see that the MH1 is unable to translate the word 'वलसाडच्या' (*valsadchya*), whereas MH9 has correctly translated it into 'वलसाड के' (*valsad ke* – *o*f Valsad) as expected in the reference translation.

5.2 10-Fold Cross Validation

To correctly compare the performance of the systems, we also did 10-fold cross validation. Results for the same are available in Table 10. We can see that significant BLEU increment in all folds of MH5 which makes use of splitting, is consistent in comparison to MH1. Also we can infer that multi-level (MH8) splitting is slightly better than two-level (MH7) and one-level (MH5) splitting.

5.3 Error Analysis

In the following subsections we analyze different errors in splitting.

5.3.1 Superfluous Splitting

With the splitting, Marathi word 'दिलावर' (*dilawar*) is getting split into 'दिला' (*dila*) + 'वर' (*war*) which is a wrong split. 'दिलावर' (*dilawar*) is a proper noun and hence should not have been split. We tried to overcome this error using NNP POS tag constraint, but that was stopping many other valid candidates from splitting. Many words like 'राजस्थानात' (*rajasthanat* – in Rajasthan) have NNP as POS tag and still are valid candidates for splitting; applying NNP POS constraint prohibits them from being split, which doesn't help in reducing sparsity in training.

5.3.2 Bad Split

Word like 'जर्मनीतील' (*jarmanitil*) is getting split into 'जर्मनीत' *(jarmanit)* + ' ील ' (*il*) which actually should have been split into 'जर्मनी' *(jarmani)* +'तील' (*til*). Similarly many words on splitting aren't giving any valid word which also doesn't help in reducing sparsity in training.

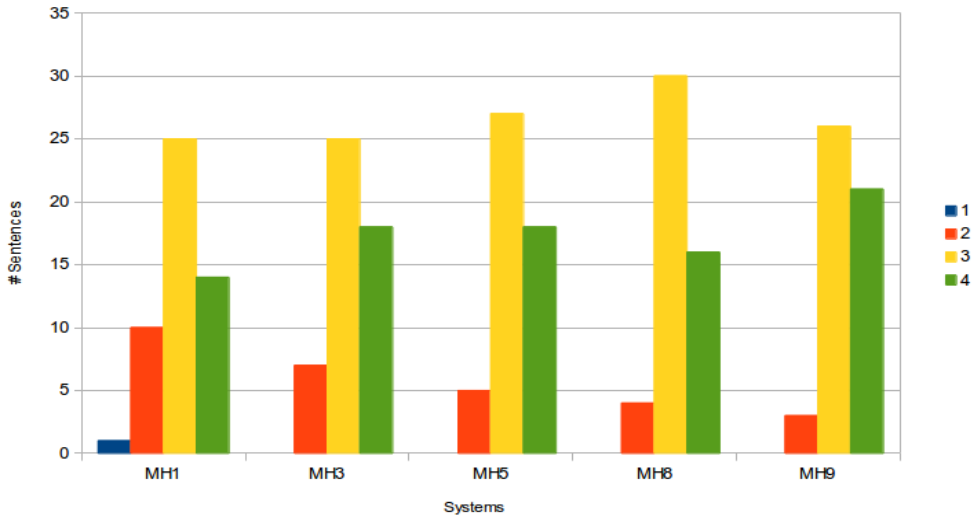

Figure 2. Manual Evaluation

	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	Avg
MH1	36.74	36.43	36.56	36.50	34.60	36.45	36.04	35.44	38.02	35.97	**36.26**
MH5	40.61	40.47	40.84	41.19	39.17	40.97	40.19	39.77	42.06	40.74	**40.60**
MH7	40.50	40.48	40.59	41.29	39.18	41.07	40.50	39.90	42.21	40.80	**40.65**
MH8	40.62	40.88	40.86	41.25	39.40	41.35	40.69	40.26	42.69	41.15	**40.90**

Table 10. 10-fold Cross Validation

Input	वलसाडच्या समुद्र किनाऱ्यावर तिथल आणि उभराट अशी सुंदर नगरे आहेत. ***valsadchya*** *samudra kinaryavar tithal ani ubhrat ashi sundar nagare aahet.*
Split Input	वलसाड च्या समुद्र किनाऱ्या वर तिथल आणि उभराट अशी सुंदर नगरे आहेत. ***Valsad chya*** *samudra kinarya var tithal ani ubhrat ashi sundar nagare aahet.*
MH1	वलसाडच्या समुद्र तट पर तिथल और उभराट ऐसे सुंदर नगर हैं । ***valasadchya*** *samudra tat par tithal aur ubharat aise sundar nagar hain.*
MH9	वलसाड के समुद्र तट पर तिथल और उभराट ऐसे सुंदर नगर हैं । ***valasad ke*** *samudra tat par tithal aur ubharat aise sundar nagar hain.*
Reference	वलसाड के समुद्र किनारे तिथल और उभराट जैसे सुंदर नगर हैं । ***valasad ke*** *samudra kinare par tithal aur ubharat jaise sundar nagar hain.*

Table 11. Comparison of Translation Systems

5.4 Comparative study with Similar Work

Not much work has been done for Marathi to Hindi Machine Translation and we compare our work with the existing systems in our knowledge. We found that the proposed system outperforms all the existing systems (Kunchukuttan et al., 2014; Shreelekha et al., 2013 and Bhosale et al., 2011). Table 13 details scores for the systems to be compared. To compare the manual evaluation we have used formula given in Figure 3 (Bhosale et al., 2011).

Accuracy= (1*N4+0.8*N3+0.6*N2)/N

N4: Number of score 4 sentences
N3: Number of score3 sentences
N2: Number of score 2 sentences
N: Total Number of sentences

Figure 3. Formula for Calculating Manual Accuracy

	BLEU	**Accuracy**
Bhosale et al., 2011	-	63.45%
Shreelekha et al., 2003	9.31	69.60%
Kunchukuttan et al., 2014	41.66	-
MH9	**42.06**	**87.20%**

Table 13. Comparison with Existing Work

6 Conclusion and Future Work

In this paper, we presented a factored Marathi to Hindi SMT system, which makes use of source side splitting and shows significantly higher accuracy than the baseline. More work remains to be done next to further take advantage of splitting by using sophisticated methodologies for the same. For example, suffix list can be enriched to include more suffixes, complex constraints can be applied to reduce negative impact of splitting, source language dictionary can be used to guide splitting and *sandhi* correction can also be exploited to generate valid words out of splitting. The same approach can be applied to other language pairs with similarities to Marathi and Hindi. SMT for Dravidian languages to Hindi is planned to be considered next.

Acknowledgments

We would like to thank the Technology Development for Indian Languages (TDIL) program and the Department of Electronics & Information Technology, Govt. of India for providing the ILCI corpus.

References

Bhosale, G., Kembhavi, S., Amberkar, A., Mhatre, S., Popale, L., & Bhattacharyya, P. (2011). Processing of Kridanta (Participle) in Marathi. In *Proceedings of ICON-2011: 9th International Conference on Natural Language Processing*. Macmillan Publishers, December 2011.

Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer R. L. & Roossin, P. S. (1990). A statistical approach to machine translation. *Computational linguistics*, *16*(2), 79-85.

Chen, S. F., & Goodman, J. (1996, June). An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics* (pp. 310-318). Association for Computational Linguistics.

Dabre, R., Amberkar, A., & Bhattacharyya, P. (2012). Morphological Analyzer for Affix Stacking Languages: A Case Study of Marathi. In *COLING (Posters)* (pp. 225-234).

Doddington, G. (2002, March). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research* (pp. 138-145). Morgan Kaufmann Publishers Inc.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume* 1:pp. 160- 167.

Heafield, K. (2011, July). KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation* (pp. 187-197). Association for Computational Linguistics.

Junghare, I. Y. (2009). Syntactic Convergence: Marathi and Dravidian. *Bulletin of the Transilvania University of Braşov• Vol*, *2*, 51.

Koehn, P., & Knight, K. (2003, April). Empirical methods for compound splitting. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1* (pp. 187-193). Association for Computational Linguistics.

Koehn, P., Och, F. J., & Marcu, D. (2003, May). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1* (pp. 48-54). Association for Computational Linguistics.

Koehn, P., & Hoang, H. (2007, June). Factored Translation Models. In *EMNLP-CoNLL* (pp. 868-876).

Kunchukuttan, A., Mishra, A., Chatterjee, R., Shah, R., & Bhattacharyya, P. (2014). Sata-Anuvadak: Tackling Multiway Translation of Indian Languages. *Pan*, *841*(54,570), 4-135.

Marcu, D., & Wong, W. (2002, July). A phrase-based, joint probability model for statistical machine translation. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10* (pp. 133-139). Association for Computational Linguistics.

Nießen, S., Och, F. J., Leusch, G., & Ney, H. (2000, May). An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *LREC*.

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*, *29*(1), 19-51.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 311-318). Association for Computational Linguistics.

Popovic, M., & Ney, H. (2004, May). Towards the Use of Word Stems and Suffixes for Statistical Machine Translation. In *LREC*.

Popović, M., Stein, D., & Ney, H. (2006). Statistical machine translation of German compound words. In *Advances in Natural Language Processing* (pp. 616-624). Springer Berlin Heidelberg.

Ramanathan, A., & Rao, D. D. (2003, April). A lightweight stemmer for Hindi. In *the Proceedings of EACL*.

Sreelekha, S. , Dabre, R., & Bhattacharyya, P. Comparison of SMT and RBMT, The Requirement of Hybridization for Marathi–Hindi MT.

Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., & Sawaf, H. (1997, September). Accelerated DP based search for statistical translation. In *Eurospeech*.