Source-Side Suffix Stripping for Bengali-to-English SMT

Rejwanul Haque, Sergio Penkale, Jie Jiang, Andy Way
Applied Language Solutions
Delph, OL3 5FZ, UK
{*firstname.lastname*}*@appliedlanguage.com*

***Abstract*—Data sparseness is a well-known problem for statistical machine translation (SMT) when morphologically rich and highly inflected languages are involved. This problem become worse in resource-scarce scenarios where sufficient parallel corpora are not available for model training. Recent research has shown that morphological segmentation can be employed on either side of the translation pair to reduce data sparsity. In this work, we consider a highly inflected Indian language as the source-side of the translation pair, Bengali. This paper presents study of morphological segmentation in SMT with a less explored translation pair, Bengali-to-English. We worked with a tiny training set available for this language-pair. We employ a simple suffix-stripping method for lemmatizing inflected Bengali words. We show that our morphological suffix separation process significantly reduces data sparseness. We also show that an SMT model trained on suffix-stripped (source) training data significantly outperforms the state-of-the-art phrase-based SMT (PB-SMT) baseline.**

***Keywords*-statistical machine translation, morphological segmentation;**

I. Introduction

The state-of-the-art phrase-based statistical machine translation (PB-SMT) model [5] does not take into account morphological information of source or target languages. The translation model in PB-SMT is learned from a considerable amount of parallel sentences. This learning process completely ignores morphological variations of source or target language tokens. A range of research including [9], [2], [10] has suggested the explicit incorporation of morphological information into SMT models in order to improve MT quality.

Indian languages are morphologically very rich and possess very large lexical variety, which results in data sparseness. As far as Indian languages are concerned, large amounts of parallel corpora are not available. [9] presented a method for translation from morphologically rich European languages to English. Like [9], in this work we apply morphological normalization techniques to a highly inflected Indian language (Bengali) for translation into English. Our morphological suffix separation process significantly reduces the data sparseness problem in the training corpus. We carried out a series of experiments with a tiny training set available for this language pair. We found that our suffix-stripped system produces statistically significant BLEU [8] improvements over the PB-SMT baseline.

The remainder of the paper is organized as follows. In Section II we discuss our motivation behind this work with an example. Section III provides an overview on resources used in our experiments and discusses our suffix-stripping method. Section IV presents the results obtained, and offers a brief qualitative analysis. In Section V we formulate our conclusions, and offer some avenues for further work.

II. Motivation

In this work, we study morphological segmentation in SMT considering a less-explored translation pair, Bengali-to-English. Bengali is an eastern Indo–Aryan language. With nearly 300 million total speakers, Bengali is the sixth most spoken language in the world. Exploring such an important language will, we hope, advance SMT research on Indian languages.

This section presents a motivating example showing how suffix stripping of Bengali inflected words reduces data sparseness. The left hand side of Figure 1 shows an inflected Bengali word *'Kolkatate'* whose literal English translation is a two-word phrase 'in Kolkata'. We also see the word *'Kolkatate'* is aligned with both 'in' and

'Kolkata'. In particular, the word *'Kolkatate'* is composed of a proper noun *'Kolkata'* (a place name) and a suffix *'te'*. Bengali nouns or pronouns are usually combined with a range of suffixes (e.g. *'ai', 'r', 'ke', 'te', 'ta'*) depending upon case and number, which results in complex inflected forms. Moreover, Bengali verbs are highly inflected and variant for tense, person, mood and aspect. Thus, rich nominal morphology and verbal inflections cause data sparseness in Bengali.

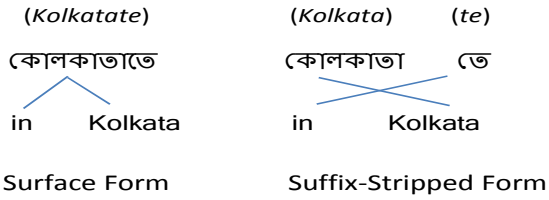

Figure 1. Example showing Bengali to English word alignments with surface form (left) and suffix-stripped form (right).

Our suffix stripping algorithm breaks up the inflected Bengali words into two parts: stem and an ending (i.e. suffix). As we see in the right hand side of Figure 1, the inflected Bengali word *'Kolkatate'* is broken up into a stem *'Kolkata'* and a suffix *'te'*. The right side of Figure 1 also shows that the stem *'Kolkata'* is aligned with the English word 'Kolkata'; and the suffix *'te'* is aligned with the English preposition 'in'. Here, the suffix *'te'* acts as a post-position whose English translation is the preposition 'in'. Thus, normalization of inflectional morphology into the base form minimizes lexical variation, which thereby counters the problem of data sparsity.

III. Data and Experimental Set-Up

A. Data

This section provides a brief overview of the data used in our experiments. As mentioned earlier, we made use of a tiny parallel corpus available for the Bengali-to-English language-pair. This parallel corpus has been taken from the EILMT[1] project. This is a tourism domain corpus. In Table I, we show the statistics of this data set with the number of sentences, source (S) and target (T) vocabulary size, number of running words, and average sentence length.

[1]English-to-Indian Language Machine Translation (EILMT) is a Ministry of IT, Govt. of India sponsored project.

Sentences	14,522	
Vocabulary Size	47,166 (S)	37,970 (T)
Running Word	294,111 (S)	344,821 (T)
Average Sentence Length	20.25 (S)	23.74 (T)

Table I
Corpus Statistics. S: Source (Bengali), T: Target (English)

In SMT, different tuning and evaluation scores may be obtained each time when we do multiple training. In particular, different results are seen when running GIZA++ [7] and MERT [6] training. This risk may be increased while working with a tiny training set. Therefore, we carried out a set of experiments to validate whether source-side suffix stripping of a highly inflected Indian language can really improve performance compared with a baseline SMT system. We conducted a total of 10 experiments (cf. Section IV) on different training, development and test sets that were randomly selected from the full data set (14,522 sentences; cf. Table I). In each experiments, training, development and test texts contain 13,022, 500 and 1,000 sentences, respectively.

Moreover, we manually created a list of 292 suffixes that were used to split inflected Bengali words. In addition to the source sentences of the parallel corpus, a large Bengali monolingual corpus from a mix of domains was used in order to enhance lemmatization quality. The combined Bengali monolingual corpus contains total of 84K sentences (with 605,932 running and 64,150 distinct words).

B. Suffix-Separation

In this section we describe a simple suffix-separation algorithm that we adopted in our experiments. To date, a small number of research has suggested novel methods to normalize inflectional Bengali words. [1] proposed an unsupervised word segmentation method into morphemes for Bengali. [10] implemented a simple list lookup approach to perform morphological split. However, we followed a simpler idea originally proposed by [3] for separating suffixes

from inflected surface words. Our simple rule-based suffix-separation method uses a list of 292 suffixes and a list of 64,150 Bengali vocabulary items (cf. Section III-A). For each word appearing in the Bengali corpus, we apply the following two steps:

1. Do longest suffix match first
2. Then, separate suffix if and only if root
   i. contains at least two characters, and
   ii. appears in the Bengali vocabulary list

We looked at the first 20 Bengali sentences (with 452 words) from the development set to measure our lemmatization accuracy. This text contains total 107 inflected words. We found that 96 inflected words were lemmatized correctly and 11 inflected words were lemmatized incorrectly. We also observed that our algorithm segmented 20 uninflected words, which is not desirable. Despite the noise inherent in our suffix stripping algorithm, we see in the next section that using this technique nevertheless brings about considerable improvements in translation quality.

IV. Evaluation Results and Discussion

This section reports experimental results obtained employing suffix-stripping techniques on our Bengali-to-English translation task. As discussed in Section III-A, we carried out a total of 10 different experiments with different training, development and test set sentences. In order to carry out our experiments, we used publicly available PB-SMT [5] toolkit Moses.[2] We evaluate our MT systems with the most widely used automatic evaluation metric: BLEU [8]. Additionally we performed statistical significance tests using bootstrap resampling [4] on BLEU. The confidence level (%) of the improvements obtained by the suffix-stripped systems with respect to the PB-SMT baseline are reported.

Table II displays BLEU scores obtained on development and test sets comparing our source-side suffix-stripped (SS) systems with the respective PB-SMT baselines. We report BLEU scores of those SS systems that give maximum and minimum gains over the PB-SMT baselines according to the evaluation scores obtained on the test sets.

[2]http://www.statmt.org/moses/

Exp	System	Dev set	Test set
Max Gain	Baseline	11.59	10.14
	SS	13.67 (100%)	13.1 (100%)
Min Gain	Baseline	11.35	12.15
	SS	12.39 (100%)	12.95 (99.8%)
Average	Baseline	11.93	11.49
	SS	13.39	12.81

Table II
Evaluation Results.

The first row of Table II shows BLEU scores of the best-performing SS system which produces 2.08 BLEU point (17.94% relative increase) and 1.87 BLEU point (18.44% relative increase) gains over the baseline on the development and test sets, respectively. The second row of Table II reports BLEU scores of the worst-performing SS system which produces 1.04 BLEU point (9.16% relative increase) and 0.80 BLEU point (6.58% relative increase) gains over the baseline on the development and test sets, respectively. We can see from Table II that all the improvements are statistically significant with respect to the baselines. The last row of Table II shows average accuracies (i.e. BLEU scores) of the SS system and the PB-SMT baseline system over the all (10) experiments. According to the average scores, the SS system gains 1.48 BLEU points (12.4% relative increase) and 1.32 BLEU points (11.48% relative increase) over the baseline on the development and test sets, respectively. This results clearly show that source-side suffix stripping is helpful in PB-SMT when translating from a highly inflected Indic language (i.e. Bengali) to English.

		Lexical Coverage (%)	
Exp	System	Development set	Test set
Max Gain	Baseline	77.88	73.79
	SS	80.94	76.31
Min Gain	Baseline	77.57	73.70
	SS	80.32	76.88
Average	Baseline	78.11	73.44
	SS	80.91	75.81

Table III
Lexical coverage (%) of test and development sets against training set.

Furthermore, we carried out an analysis to see how source-language morphological analysis helped in improving MT quality. Our intention

behind performing morphological segmentation was to reduce sparseness in the Bengali corpus. We calculated the lexical coverage of the development set and test set against the training set twice (i.e. before and after suffix separation) in order to see whether our segmentation process can reduce the sparseness of the training set. Table III shows lexical coverage (%) of the development and test sets against the respective training set for each of the experiments reported in Table II. The last row of Table III shows average scores (i.e. lexical coverage) of the development and test sets over the all (10) experiments. We clearly see from Table III that morphological segmentation brings a 2–3% increase in lexical coverage of the development and test sets against the training set. In other words, our suffix-separation program reduces data-sparseness and improves the lexical coverage of test set sentences, which thereby helps in improving MT quality.

V. Conclusions and Future Work

In this work we introduce a less explored language-pair to study SMT: Bengali-to-English. Due to the morphological richness and highly inflectional behavior of the Bengali language, the tiny training corpus available for this language-pair was extensively sparse. In this work we employ a suffix-stripping algorithm on inflected Bengali words to split them into their base form and suffixes, which in turn reduces data sparseness. Then, we carried out a set of experiments by randomly choosing training, development and test sets from parallel corpus to make up for the fact that we had only a small amount of parallel data. We found that each of our source-side suffix-stripped systems significantly outperform the respective PB-SMT baseline according to the MT evaluation scores obtained.

A common property of all Indian languages (Indo-Aryan: Hindi, Marathi, Punjabi, etc. and Dravidian: Tamil, Telugu, Malayalam, etc.) including Bengali is that these languages are morphologically very rich and highly inflected. Therefore, the technique proposed in this paper to improve Bengali-to-English SMT is also applicable to the other Indian languages.

In future, we would like to study SMT by applying our approach to other Indian languages. We also intend to perform a deep manual qualitative analysis on the MT output to compare our suffix-stripped systems with the PB-SMT baseline.

References

[1] Dasgupta, S. and V. Ng. 2004. Unsupervised Morphological Parsing of Bengali. *Language Resources and Evaluation,* 3–4:311–330.

[2] Goldwater, S. and D. McClosky. 2005. Improving statistical MT through morphological analysis. In *HLT-EMNLP-2005: Proceedings of Human Technology Conference and Conference on Empirical Methods in Natural Language Processing* pages 676–683, Vancouver.

[3] Keshava, S. and E. Pitler. A simpler, intuitive approach to morpheme induction. In *PASCAL challenge workshop on unsupervised segmentation of words into morphemes*, Venice, Italy, 2006.

[4] Koehn, P. Statistical significance tests for machine translation evaluation. In *EMNLP-2004: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain, 2004.

[5] Koehn, P, F. J. Och, and D. Marcu. Statistical Phrase-Based Translation. In *HLT-NAACL 2003: conference combining Human Language Technology conference series and the North American Chapter of the Association for Computational Linguistics conference series*, pages 48–54, Edmonton, AB, 2003.

[6] Och, F. J. Minimum Error Rate Training in Statistical Machine Translation. In *41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, pages 160–167, Sapporo, Japan, 2003.

[7] Och, F. J. and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

[8] Papineni, K., S. Roukos, and W. Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, pages 311–318, Philadelphia, PA, 2002.

[9] Popović, M., and H. Ney. 2004. Towards the use of word stems and suffixes for statistical machine translation. In *LREC-2004: Fourth International Conference on Language Resources and Evaluation, Proceedings,* pages 1585–1588, Lisbon, Portugal.

[10] Roy, M. *Approches to handle scare resources for Bengali statistical machine translation*. PhD Thesis, Simon Fraser University, BC, Canada, 2010.