Abstract

This paper presents Centre for Development of Advanced Computing Mumbai’s (CDACM) submission to NLP Tools Contest on Statistical Machine Translation in Indian Languages (ILSMT) 2015 (collocated with ICON 2015). The aim of the contest is to collectively explore the effectiveness of Statistical Machine Translation (SMT) while translating within Indian languages and between English and Indian languages.

In this paper, we report our work on all five pairs of languages, namely Bengali-Hindi, Marathi-Hindi, Tamil-Hindi, Telugu-Hindi and English-Hindi for Health, Tourism and General domains. We have used suffix separation, compound splitting and pre-reordering prior to SMT training and testing.

1 Introduction

In this paper, we present our SMT experiments from Bengali, Marathi, Tamil, Telugu and English to Hindi. From the set of languages involved, Bengali, Hindi and Marathi belong to the Indo-Aryan family and Tamil and Telugu are from Dravidian language family. All languages except English, have the same flexibility towards word order, canonically following the Subject-Object-Verb (SOV) structure.

With reference to morphology Bengali, Marathi, Tamil and Telugu are more agglutinative compared to Hindi. 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 sentences before training the SMT system. To handle this morphological difference we have used suffix separation and compound word splitting (Pimpale et al., 2014; Patel et al., 2014).

For English to Hindi SMT pre-reordering developed by (Patel et al., 2013) has been used. And better alignment is achieved using stem as an alignment factor (Koehn et al., 2007).

All machine translation (MT) systems suffer from Out of Vocabulary (OOV) words. These OOV words are mostly named entities, technical terms and foreign words which can be translated using transliteration system. We have used Durrani et al. (2014), which is a fully unsupervised approach for developing a transliteration system using parallel corpus meant for SMT training.

The rest of the paper is organized as follows. In Section 2, we discuss Dataset and Experimental setup. Section 3 discusses experiments and results. Submitted systems to the shared task are described in section 4 followed by conclusion and future work in section 5.

2 Data-set and Experimental Setup

In the following subsections we describe Training and Testing corpus followed by pre-processing and SMT system setup for experiments.

2.1 Corpus for SMT Training and Testing

We have used corpus shared by ILSMT detailed in Table 1 for the experiments. Testing for the experiments were done using heldout data which is shared as the development set. For unconstrained systems additional data (Bojar et al., 2014; Khapra et al., 2010) has been used for language modeling.

|          | Health | Tourism | General |
|----------|--------|---------|---------|
| Training(TM) | 24000  | 24000   | 48000   |
| Training(LM) | 48000  | 48000   | 48000   |
| Test 1     | 500    | 500     | 1000    |

Table 1. Training, Testing and Development Data, TM - Translation Model, LM - Language Model
To tackle morphological divergence between the source and target languages for the purpose of a better SMT system, we preprocessed the source (Bengali, Marathi, Tamil and Telugu) for suffix separation and compound word splitting (Pimpale et al., 2014; Patel et al., 2014) prior to training and testing. To handle the structural divergence (English-Hindi), we have used source side reordering (Patel et al., 2013).

2.3 Transliteration (TR)

We use top n-best transliteration output for OOV words. These candidates are plugged in and re-scored with the language model to get the best translation for source sentence.

2.4 Language Modelling

Training LM with huge amount of monolingual data (approx 10GB, ) requires good computing resource (200GB storage space in working directory, 60GB RAM). In Bojar et al., 2014 corpus, we need to remove `<s>` as KENLM doesn’t support the same. We have trained 5-gram LM with modified Kneser-Ney smoothing. LM size was quite large (ARPA-88.7GB, binary-27.4GB) even after binarization.

2.5 SMT System Set Up

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 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) lightweight stemmer for Hindi. For English we have used porter stemmer (Minnen et. al., 2001).

2.6 Evaluation Metrics

The different experimental systems are being compared using, BLEU (Papineni et al., 2002), NIST (Doddington, 2002), translation error rate (TER) (Matthew, et al., 2006). For a MT system to be better, higher BLEU and NIST scores with lower TER are desired.

3 Experiments and Results

Table 2 shows bleu scores for various systems tried under constrained submission, i.e. systems trained only on the shared data. We can see, the use of the preprocessing and transliteration has contributed to the improvement of around 1-5 bleu points across

|                         | Bengali-Hindi | English-Hindi | Marathi-Hindi | Tamil-Hindi | Telugu-Hindi |
|-------------------------|--------------|--------------|--------------|------------|--------------|
|                         | S1 | S2 | S3 | S1 | S2 | S3 | S1 | S2 | S3 | S1 | S2 | S3 |
| General BLEU            | 30.18 | 31.23 | **32.22** | 18.76 | 23.15 | **23.44** | 35.74 | 40.01 | **40.21** | 16.64 | 20.17 | **20.58** |
| NIST                    | 6.888 | 6.943 | **7.092** | 5.862 | 6.037 | **6.092** | 7.511 | 7.805 | **7.867** | 4.742 | 5.303 | **5.391** |
| TER                     | 48.43 | 47.27 | **46.46** | 64.05 | 59.89 | **59.58** | 42.75 | 39.95 | **39.62** | 64.68 | 62.65 | **62.12** |

Table 2. Bleu scores for different experimental systems (CONSTRAINED).

|                         | Bengali-Hindi | English-Hindi | Marathi-Hindi | Tamil-Hindi | Telugu-Hindi |
|-------------------------|--------------|--------------|--------------|------------|--------------|
|                         | S1 | S2' | S3' | S1 | S2' | S3' | S1 | S2' | S3' | S1 | S2' | S3' |
| General BLEU            | 30.18 | 31.58 | **33.77** | 18.76 | 19.96 | **24.00** | 35.74 | 36.80 | **41.20** | 16.64 | 16.38 | **20.38** |
| NIST                    | 6.888 | 7.080 | **7.195** | 5.862 | 5.922 | **6.121** | 7.511 | 7.617 | **7.935** | 4.742 | 4.677 | **5.340** |
| TER                     | 48.43 | 47.03 | **45.52** | 64.05 | 63.03 | **58.99** | 42.75 | 42.35 | **39.25** | 64.68 | 64.45 | **62.25** |

Table 3. Bleu scores for different experimental systems (UNCONSTRAINED).

S1- BL, S2- BL+PP, S3- BL+PP+TR
BL-Baseline, TR- Transliteration

S1- BL, S2'- BL+ELM, S2'- BL+PP+ELM
BL-Baseline, ELM- Extended LM (Additional data used for LM training)
the language pairs. Detailed evaluation scores for unconstrained systems, i.e. systems using language model built on external data sources, are in table 3. Significant improvement can be observed from the table when additional data have been used for language modeling.

### 4 Submission to the Shared Task

We have submitted two different result sets for the contest, namely constrained and unconstrained. The constrained systems were trained only on the data shared by the organizers. For unconstrained systems, we have used additional monolingual data which include Bojar et al. (2014) and Khapra et al. (2010), for language modeling. Table 4 and 5 summarizes the evaluation of the two systems submitted to the shared task.

### 5 Conclusion and Feature Work

In this paper, we presented various systems for translation from Bengali, English, Marathi, Tamil and Telugu to Hindi. These SMT systems with the use of source side suffix separation, compound splitting and pre-reordering show significantly higher accuracy over the baseline. More work remains to be done next to further improve the experimented pre-processing techniques for better translation quality.

| Language Pair      | Health BLEU | Health FSS | Tourism BLEU | Tourism FSS | General BLEU | General FSS |
|--------------------|-------------|------------|--------------|-------------|--------------|-------------|
| Bengali-Hindi      | 30.93       | 32.95      | 28.91        | 32.05       | 30.18        | 46.46       |
| English-Hindi      | 20.26       | 23.39      | 16.80        | 21.41       | 18.76        | 60.58       |
| Marathi-Hindi      | 35.74       | 39.31      | 35.34        | 39.87       | 42.75        | 59.58       |
| Tamil-Hindi        | 17.87       | 21.19      | 15.48        | 19.34       | 16.64        | 25.12       |
| Telugu-Hindi       | 24.90       | 30.21      | 20.58        | 24.88       | 28.58        | 52.18       |

Table 4. Consolidated Results, FSS- Final Submitted System, (CONSTRAINED)

| Language Pair      | Health BLEU | Health FSS | Tourism BLEU | Tourism FSS | General BLEU | General FSS |
|--------------------|-------------|------------|--------------|-------------|--------------|-------------|
| Bengali-Hindi      | 30.93       | 33.45      | 28.91        | 32.05       | 30.18        | 46.46       |
| English-Hindi      | 20.26       | 24.18      | 16.80        | 21.41       | 18.76        | 60.58       |
| Marathi-Hindi      | 35.74       | 39.31      | 35.34        | 39.87       | 42.75        | 59.58       |
| Tamil-Hindi        | 17.87       | 21.19      | 15.48        | 19.34       | 16.64        | 25.12       |
| Telugu-Hindi       | 24.90       | 30.21      | 20.58        | 24.88       | 28.58        | 52.18       |

Table 5. Consolidated Results, FSS- Final Submitted System, (CONSTRAINED)

### References

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.

Bojar, Ondřej; Diatka, Vojtěch; Rychlí, Pavel; Straňák,
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.

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.

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.

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).

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.

Minnen, Guido, John Carroll, and Darren Pearce. Applied morphological processing of English. Natural Language Engineering 7.03 (2001): 207-223.

Khapra, M. M., Kulkarni, A., Sohoney, S., & Bhattacharyya, P. (2010, July). All words domain adapted WSD: Finding a middle ground between supervision and unsupervision. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (pp. 1532-1541). Association for Computational Linguistics.

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

Patel, R. N., Gupta, R., Pimpale, P. B. & Sasikumar, M. (2013, August). Reordering rules for English-Hindi SMT. In Proceedings of the Second Workshop on Hybrid Approaches to Translation (pp. 34-41). Association for Computational Linguistics.

Patel, R. N., Pimpale, P. B., & Sasikumar, M. (2014, December). Statistical Machine Translation for Indian Languages: Mission Hindi. In Proceedings of NLP Tools contest, ICON-2014: 11th International Conference on Natural Language Processing.