IITP English-Hindi Machine Translation System at WAT 2016

Sukanta Sen, Debajyoty Banik, Asif Ekbal, Pushpak Bhattacharyya
Department of Computer Science and Engineering
Indian Institute of Technology Patna
Bihar, India
{sukanta.pcs15,debajyoty.pcs13,asif,pb}@iitp.ac.in

Abstract

In this paper we describe the system that we develop as part of our participation in WAT 2016. We develop a system based on hierarchical phrase-based SMT for English to Hindi language pair. We perform reordering and augment bilingual dictionary to improve the performance. As a baseline we use a phrase-based SMT model. The MT models are fine-tuned on the development set, and the best configurations are used to report the evaluation on the test set. Experiments show the BLEU of 13.71 on the benchmark test data. This is better compared to the official baseline BLEU score of 10.79.

1 Introduction

In this paper, we describe the system that we develop as part of our participation in the Workshop on Asian Translation (WAT) 2016 (Nakazawa et al., 2016) for English-Hindi language pair. This year English-Hindi language pair is adopted for translation task for the first time in WAT. Apart from that, the said language pair was introduced in WMT 14 (Bojar et al., 2014). Our system is based on Statistical Machine Translation (SMT) approach. The shared task organizers provide English-Hindi parallel corpus for training and tuning and monolingual corpus for building language model. Literature shows that there exists many SMT based appraoches for differnt language pairs and domains. Linguistic-knowledge independent techniques such as phrase-based SMT (Koehn et al., 2003) and hierarchical phrase-based SMT (Chiang, 2005; Chiang, 2007) manage to perform efficiently as long as sufficient parallel text are available. Our submitted system is based on hierarchical SMT, performance of which is improved by performing reordering in the source side and augmenting English-Hindi bilingual dictionary.

The rest of the paper is organized as follows. Section 2 describes the various methods that we use. Section 3 presents the details of datasets, experimental setup, results and analysis. Finally, Section 4 concludes the paper.

2 Method

For WAT-2016, we have submitted two systems for English to Hindi (En-Hi) translation, *viz.* one without adding any external data to the training corpus and the other by augmenting bilingual dictionary in training. Both systems are reordered in the source side. As a baseline model we develop a phrase-based SMT model using Moses (Koehn et al., 2007). We perform several experiments with the hierarchical SMT in order to study the effectiveness of reordering and bilingual dictionary augmentation. These were done to improve syntactic order and alignment with linguistic knowledge.

2.1 Phrase-based Machine Translation

Phrase-based statistical machine translation (PBSMT) (Koehn et al., 2003) is the most popular approach among all other approaches to machine translation and it has became benchmark for machine translation systems in academia as well as in industry. A phrase-based SMT consists

Proceedings of the 3rd Workshop on Asian Translation,
pages 216–222, Osaka, Japan, December 11-17 2016.

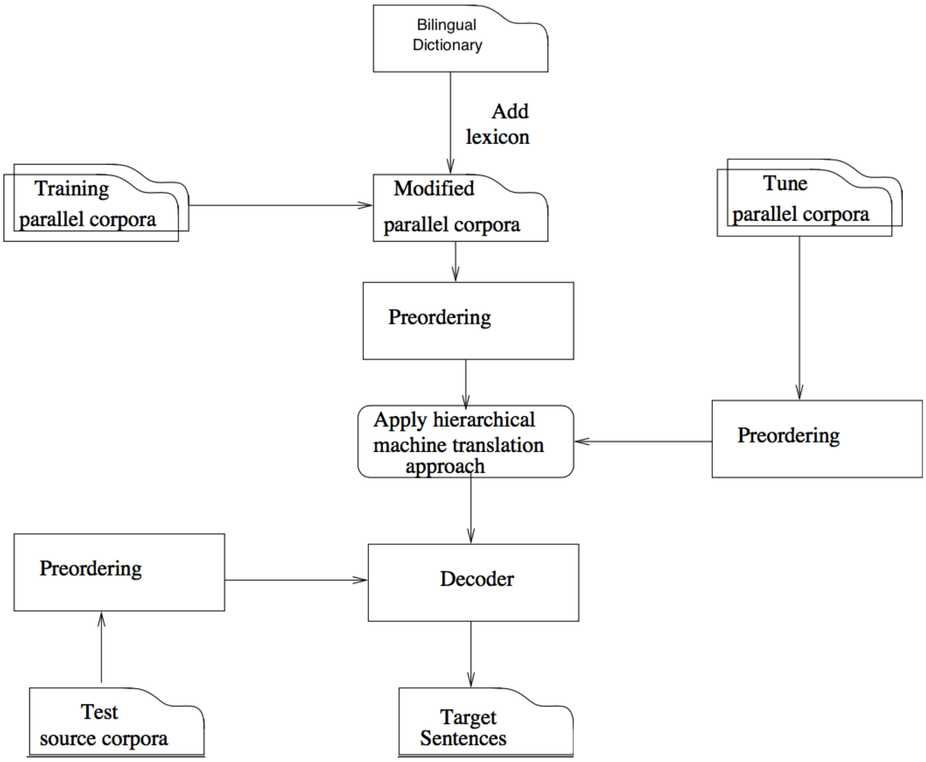

Figure 1: Hierarchical approach with reordering and dictionary augmentation.

of a language model, a translation model and a distortion model. Mathematically, it can be expressed as:

$$e_{best} = argmax_e P(e|f) = argmax_e [P(f|e) P_{LM}(e)] \tag{1}$$

where, e_{best} is the best translation, f is the source sentence, e is target sentence, $P(f|e)$ and $P_{LM}(e)$ are translation model and language model respectively. $P(f|e)$ (translation model) is further decomposed in phrase based SMT as,

$$P(\bar{f_1}^I|\bar{e_1}^I) = \prod_{i=1}^{I} \phi(\bar{f_i}|\bar{e_i}) d(start_i - end_{i-1} - 1)$$

where, $\phi(\bar{f_i}|\bar{e_i})$ is, the probability that the phrase $\bar{f_i}$ is the translation of the phrase $\bar{e_i}$, known as phrase translation probability which is learned from parallel corpus and $d(start_i - end_{i-1} - 1)$ is distortion probability which imposes an exponential cost on number of input words the decoder skips to generate next output phrase. Decoding process works by segmenting the input sentence f into sequence of I phrases $\bar{f_1}^I$ distributed uniformly over all possible segmentations. Finally, it uses beam search to find the best translation.

2.2 Hierarchical Phrase based model

Phrase-based model treats phrases as atomic units, where a phrase, unlike a linguistic phrase, is a sequence of tokens. These phrases are translated and reordered using a reordering model to produce an output sentence. This method can robustly translate the phrases that commonly occur in the training corpus. But authors (Koehn et al., 2003) have found that phrases longer than three words do not improve the performance much because data may be too sparse for learning longer phrases. Also, though the phrase-based approach is good at reordering of words but it fails at long distance phrase reordering using reordering model. To over come this, (Chiang, 2005) came up with Hierarchical PBSMT model which does not interfere with the strengths of the PBSMT instead capitalizes on them. Unlike the phrase-based SMT, it uses hierarchical phrases that contain sub-phrases. A weighted synchronous context-free grammar is induced from the parallel corpus and the weight of a rule tells the decoder how probable the rule is. Decoder implements a parsing algorithm which is inspired by monolingual syntactic chart parsing along with a beam search to find the best target sentence.

2.3 Reordering

One of the major difficulties of machine translation lies in handling the structural differences between the language pair. Translation from one language to another becomes more challenging when the language pair follows different word order. For example, English language follows subject-verb-object (SVO) whereas Hindi follows subject-object-verb (SOV) order. Research (Collins et al., 2005; Ramanathan et al., 2008) has shown that syntactic reordering of the source-side to conform the syntax of the target-side alleviates the structural divergence and improves the translation quality significantly. Though the PBSMT has an independent reordering model which reorders the phrases but it has limited potential to model the word-order differences between different languages (Collins et al., 2005).

We perform syntactic reordering of the source sentences in the preprocessing phase in which every English sentence is modified in such a way that its word order is almost similar to the word order of the Hindi sentence.
For example,

English: The president of America visited India in June.

Reordered: America of the president June in India visited.

Hindi: अमेरिका के राष्ट्रपति ने जून में भारत की यात्रा की।
(amerikA ke rAShTrapati ne jUna meM bhArata kI yAtrA kI .)

For source-side reordering we use the rule-based preordering tool[1] , which takes parsed English sentence as input and generates sentence whose word order is similar to that of Hindi. This reordering is based on the approach developed by (Patel et al., 2013) which is an extension of an earlier work reported in (Ramanathan et al., 2008). For parsing source side English sentences, we use Stanford parser[2].

2.4 Augmenting Bilingual Dictionary

Bilingual dictionaries are always useful in SMT as it improves the word-alignment which is the heart of every SMT. In addition to reordering the source corpus, we add a English-Hindi bilingual dictionary to improve our MT system. We show our proposed model in Figure 1. We use Moses (Koehn et al., 2007), an open source toolkit for training different systems. We start training with Phrase-based SMT as a baseline system. Then, augment bilingual dictionary to the training corpus and perform reordering in the source side to improve syntactic order. Thereafter, we train a hierarchical phrase-based SMT model. For preparing bilingual dictionary, we use English-Hindi bilingual mapping[3] which contains many Hindi translations for each English word. We preprocess it and add it to the parallel corpus. After preprocessing, it contains 157975 English-Hindi word translation pairs.

2.5 Data Set

For English-Hindi task, we use IIT Bombay English-Hindi Corpus[4] which contains training set, test set, development set and as well as a monolingual Hindi corpus. The training set was collected from the various existing sources. However, development set and test set are the same

[1] http://www.cfilt.iitb.ac.in/ moses/download/cfilt_preorder

[2] http://nlp.stanford.edu/software/lex-parser.html

[3] http://www.cfilt.iitb.ac.in/ sudha/bilingual_mapping.tar.gz

[4] http://www.cfilt.iitb.ac.in/iitb_parallel/

newswire test and development set of WMT 14. The corpus belongs to miscellaneous domain. Train set consists of 1,492,827 parallel sentences, whereas test set and development set contain 2,507 and 520 parallel sentences, respectively. Monolingual Hindi corpus comprises 45,075,279 sentences. Table 1 shows the details of the corpus.

Set	#Sentences	#Tokens	
		En	Hi
Train	1,492,827	20,666,365	22164816
Test	2507	49,394	57,037
Development	520	10,656	10174
Monolingual Hindi corpus	45,075,279	844,925,569	

Table 1: Statistics of data set

2.6 Preprocessing

We begin with a preprocessing of raw data, which includes tokenization, true-casting, removing long sentences as well as sentences with a length mismatch exceeding certain ratio. Training and development sets were already tokenized. For tokenizing English sentences we use tokenizer.perl[5] script and for Hindi sentences we use indic_NLP_Library[6].

2.7 Training

For all the systems we train, we build n-gram (n=4) language model with modified Kneser-Ney smoothing (Kneser and Ney, 1995) using KenLM (Heafield, 2011). We build two separate language models, one using the monolingual Hindi corpus and another merging the Hindi training set with the monolingual corpus. In our experiment, as we find language model built using only monolingual Hindi corpus produces better results in terms of BLEU (Papineni et al., 2002) score therefore, we decide to use the former language model. For learning the word alignments from the parallel training corpus, we used GIZA++ (Och and Ney, 2003) with grow-diag-final-and heuristics.

We build several MT systems using Moses based on two models, namely phrase-based SMT and hierarchical phrase-based SMT. For building phrase-based systems, we use msd-bidirectional-fe as reordering model, set distortion limit to 6. For other parameters of Moses, default values were used. For building hierarchical phrase-based systems we use default values of the parameters of Moses. Finally, the trained system was tuned with Minimum Error Rate Training (MERT) (Och, 2003) to learn the weights of different parameters of the model.

3 Results and Analysis

We build the following systems using Moses[7].

1. Phrase-based model (Phr)

2. Phrase-based model after reordering the source side (PhrRe)

3. Hierarchical phrase-based model (Hie)

4. Hierarchical phrase-based model after reordering the source side. We build two variations of this model: one (HieRe) without adding any external resources to the train set and another (HieReDict) with adding bilingual dictionary to the train set.

[5] https://github.com/moses-smt/mosesdecoder/blob/RELEASE-3.0/scripts/tokenizer/tokenizer.perl

[6] https://bitbucket.org/anoopk/indic_nlp_library

[7] https://github.com/moses-smt/mosesdecoder

We evaluate each system using BLEU metric on WMT 14 test set. The official baseline model reports the BLEU of 10.79 for En-Hi translation task. Our baseline, phrase-based model using 4-gram language model achieves the BLEU score of 11.79. In our hierarchical phrase-based model we obtain the BLEU score of 13.18. After reordering the source (i.e. English corpus), we obtain the BLEU score of 13.56 in the hierarchical phrase based SMT. The performance is further improved to 13.71 when we augment English-Hindi bilingual dictionary with the training set. We summarize the BLEU scores of the different systems in Table 2.

Approach	**BLEU Score**
Baseline (official)	10.79
Phr	11.79
Hie	13.18
HieRe	13.57
HieReDict	13.71

Table 2: Results of different models

We study the the output sentences of the final system and classify the error according to the linguistic error categories as given in (Vilar et al., 2006) and find that the following are the the most common errors.
1.The translated word is not according to the context.

> **Source:** cat is a brand that has been offering strong, durable, beautiful and high quality products for the past one hundred years .
> **Reference:** कैट एक ऐसा ब्राड है जो पिछले सौ साल से मजबूत, टिकाऊ, सुंदर और बेहतरीन उत्पाद पेश कर रहा है।
> (kaiTa eka aisA brADa hai jo piChale sau sAla se majabUta, TikAU, suMdara aura behatarIna utpAda pesha kara rahA hai.)
> **Output:** बिल्ली एक ब्रांड है कि पिछले सौ वर्षों के लिए मजबूत, टिकाऊ, सुन्दर और उच्च गुणवत्ता वाले उत्पादों की पेशकश की गई है।
> (billI eka brAMDa hai ki piChale sau varShoM ke lie majabUta, TikAU, sundara aura uchcha guNavattA vAle utpAdoM kI peshakasha kI gaI hai.)

Here, the word ”cat” is translated as बिल्ली (billI) which is a wrong translation in the context of the source sentence.
2. Word order error.

> **Source:** all the guests will join the Lakshmi Puja for the birthday party on Friday.
> **Reference:** शुक्रवार को सभी मेहमान यहां जन्मदिन पर लक्ष्मी पूजा के लिए जुटेंगे।
> (shukravAra ko sabhI mehamAna yahAM janmadina para lakShmI pUjA ke lie juTeMge.)
> **Output:** सभी अतिथियों का जन्मदिन शुक्रवार को लक्ष्मी पूजन में शामिल होंगे।
> (sabhI atithiyoM kA janmadina shukravAra ko lakShmI pUjana meM shAmila hoMge.)

Here, words in the output are not properly ordered, the correct word ordering is the following: शुक्रवार को सभी अतिथियों जन्मदिन का लक्ष्मी पूजन में शामिल होंगे। (shukravAra ko sabhI atithiyoM janmadina kA lakShmI pUjana meM shAmila hoMge.).

We also find that test set contains longer sentences compared to the training set. Average sentence lengths of training sentences are approximately 14 and 15 for English and Hindi, respectively, whereas for test set, average sentence lengths are approximately 20 and 23, respectively. Now we give some examples where reordering and dictionary augmentation improve translation outputs.
1. Example 1

Source: the rain and cold wind on Wednesday night made people feel cold.
Hie: बुधवार रात को बरसात और ठंडी हवा ने लोगों को ठंड लग रही थी।
(budhavAra rAta ko barasAta aura ThaMDI havA ne logoM ko ThaMDa laga rahI thI.)
HieRe: बुधवार की रात को बरसात और ठंडी हवा से ठंडक महसूस हुई।
(budhavAra kI rAta ko barasAta aura ThaMDI havA se ThaMDaka mahasUsa huI.)
HieReDict: बुधवार रात बारिश और ठंडी हवा से लोगों को ठंड लगने लगा।
(budhavAra rAta bArisha aura ThaMDI havA se logoM ko ThaMDa lagane lagA.)

In the above example, **Hie** approach generates wrong postposition ने (ne), whereas **HieRe** outputs correct postposition से (se). So reordering helps here but it drops the word लोगों (logoM), which is brought back by **HieReDict** approach.

2. Example 2

Source: he demanded the complete abolition of house tax in Panchkula.
Hie: वे को पंचकूला में हाउस टैक्स को समाप्त करने की मांग की।
(ve ko paMchakUlA meM hAusa Taiksa ko samApta karane kI mAMga kI.)
HieRe: वह पंचकुला में हाउस टैक्स से पूरी तरह दूर करने की मांग की।
(vaha paMchakulA meM hAusa Taiksa se pUrI taraha dUra karane kI mAMga kI.)
HieReDict: पंचकूला में हाउस टैक्स की पूरी तरह समाप्ति की मांग की।
(paMchakUlA meM hAusa Taiksa kI pUrI taraha samApti kI mAMga kI.)

Here, **Hie** approach generates wrong output वे को (ve ko) for source word 'he' but reordering helps by translating it as वह (vaha). Also, we can see when we add dictionary, it generates better Hindi translation समाप्ति (samApti) for source word 'abolition'.

It is not that reordering and augmenting dictionary always helps. There are some source sentences for which these approaches deteriorate the translation quality but these two approaches improve the overall system.

4 Conclusion

In this paper we describe the system that we develop as part of our participation in the shared task of WAT 2016. We have submitted models for English-Hindi language pair. We have developed various models based on phrase-based as well as hierarchical MT models. Empirical analysis shows that we achieve the best performance with a hierarchical SMT based approach. We also show that hierarchical SMT model, when augmented with bilingual dictionary along with syntactic reordering of English sentences produces better translation score.

5 Acknowledgments

We acknowledge Salam Amitra Devi *(lexicographer, IIT Patna)* for contributing in qualitative error analysis.

References

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 263–270. Association for Computational Linguistics.

David Chiang. 2007. Hierarchical phrase-based translation. *computational linguistics*, 33(2):201–228.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 531–540. Association for Computational Linguistics.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197. Association for Computational Linguistics.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, pages 181–184. IEEE.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. 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*, pages 48–54. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

Toshiaki Nakazawa, Hideya Mino, Chenchen Ding, Isao Goto, Graham Neubig, Sadao Kurohashi, and Eiichiro Sumita. 2016. Overview of the 3rd workshop on asian translation. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, Osaka, Japan, December.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

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*, pages 160–167. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Raj Nath Patel, Rohit Gupta, Prakash B Pimpale, and Sasikumar M. 2013. Reordering rules for english-hindi smt. In *Proceedings of the Second Workshop on Hybrid Approaches to Translation*, pages 34–41. Association for Computational Linguistics.

Ananthakrishnan Ramanathan, Jayprasad Hegde, Ritesh M Shah, Pushpak Bhattacharyya, and M Sasikumar. 2008. Simple syntactic and morphological processing can help english-hindi statistical machine translation. In *IJCNLP*, pages 513–520.

David Vilar, Jia Xu, Luis Fernando d'Haro, and Hermann Ney. 2006. Error analysis of statistical machine translation output. In *Proceedings of LREC*, pages 697–702.