Does Syntactic Knowledge help English-Hindi SMT ?

Taraka Rama, Karthik Gali, Avinesh PVS
Language Technologies Research Centre,
{taraka,karthikg,avinesh}@students.iiit.ac.in

Abstract

In this paper we explore various parameter settings of the state-of-art Statistical Machine Translation system to improve the quality of the translation for a 'distant' language pair English-Hindi. We propose new techniques for improving the quality of translation. A slight improvement over the baseline is reported using these techniques. We also show that removal of redundant translations from the phrase table can improve the quality of the translation significantly.

1 Introduction

Statistical Machine Translation techniques have been quite successful for translation between European languages, in the recent years. MOSES(Koehn et al., 2007), the SMT tool kit is used very frequently in the machine translation experiments and also for deploying real time systems. A glimpse of papers being published at any major NLP conference shows that the techniques have been widely popular not only between linguistically related European languages but also between English-Arabic and English-Chinese, which are linguistically distant pairs. There has been some work (Venkatapathy and Bangalore,) on the English-Hindi Machine translation but a lot remains to be explored. In this paper we propose new reordering techniques for improving the translation quality. The translation quality is evaluated using the widely popular Bleu metric(Papineni et al., 2001). We compare ourselves with the baseline score reported by the organisers and report the improvements. The paper is structured as follows. In Section 2 we identify the problem areas, Section 3 elaborates on our approaches, Section 4 briefs about the data sets, Section 5 talks about the initial experiments with different Moses settings. Section 6 describes the various experiments performed and in Section 7, we make observations based on the experiments' results and Section 8 discusses the new avenues to be explored.

2 Scope for Improvement

We compare the Bleu score of English-Hindi with English-German, English-French in Table 1. The Bleu score for the English-French and English-German language pairs were obtained from the Statistical Machine Translation website[1]. For the English-Hindi pair, the Bleu score is reported as given in the Shared Task website[2].

The differences in the sizes of the training sets have to be kept in mind, when reporting the Bleu score for a language pair. The huge difference in the Bleu scores of English-French and English-Hindi is also due to the difference in the size of the training corpora. When the size of the training corpora are normalised, the real difference in the Bleu scores can be observed. At this point we only have a single reference translation. Ideally, the system has to be compared against the multiple refernce data sets and then Bleu score can be an average of all the Bleu scores. One obvious reason for the lower Bleu score is the linguistic distance between the two languages. The underlying structural differences between the two languages manifest themselves as a relatively low score. The structural differences can be broken into smaller subproblems and then can be addressed separately.

[1] http://www.statmt.org/matrix/
[2] http://ltrc.iiit.ac.in/nlptools2008/resources2.php

Language-Pair	Bleu Score
English-French	31.36
English-German	25.36
English-Hindi	17.70

Table 1: Comparision of Bleu Scores for Various language Pairs

We have identified the following problem areas where there is a huge scope for improvization. The most significant difference is in the word order (or chunk order) of the two languages.

Word order forms a major weighting factor for the low Bleu score. The performance of Moses is quite high, when it comes to word to word translation. This claim is corroborated by the unigram score which is as high as 60. Data sparseness manifests itself as unknown words leading to low Bleu scores. So the major thrust in this paper is to tackle the issue of reordering. Nonethless, the problem of unknown word translation is also examined.

3 Our Approach

Our hypothesis is that instead of trying to reorder the target language ouput using POS tags or huge language models or chunk language models, we instead go for the rearrangement of the words in the source language itself. The quality can also be improved by using richer and huge language models. This in itself, is a costly procedure. Especially, for a language like English which has syntactic parsers of high quality, it is always desirable to tap the existing resources. Our modelling goes as follows. The core idea is that learning the reordering relations from the parse of the source language sentence would fetch better results. The source sentence is parsed and hence reordered by learning the source-target reordering relations using POS tags on both the sides. This scheme would ensure that the transfer rules learnt are more generic and hence the intution that they would improve the performance of the system.

The second hypothesis is concerning the presence of the unknown words in the target language output. Two possible ways are factored models as given in(Koehn and Hoang, 2007) and using a bilingual dictionary to translate the lemmas and if the unknown word is a named entity then transliterate it. Recently, factored models are becoming famous and hence we opted out to test the factored models using POS taggers and morph analysers, which are available for both English and Hindi side. We toyed with the distortion limit and we observed that if we allow unlimited reordering the Bleu score improved significantly. We found no significant improvement over the old models. The hypothesis is falsified if the Bleu score shows no improvement.

4 Data Sets and Baseline

We initially explored various parameters of the MOSES to see how the quality improves. The training data for learning word alignments was conducted on 7000 sentence parallel corpora provided in the Shared Task(taken from EILMT corpus). We also went on to check how the accuracy increases by increasing the size of the monolingual corpus on which language model is trained. For the baseline the following parameters were adopted. The distortion limit was kept at 6 and the lexicalised reordering models were trained by using msd-bidirectional-fe. The distortion limit shows how often the phrases get reordered. The MOSES has both a distance based reordering model and a lexicalised reordering model. Lexicalised reordering model gives the orientation of a phrase. The distortion limit is simply the absolute of the difference of the last word of the previously translated English phrase and the position of the first word in the currently translated phrase. The language model was trained on the hindi side of the training data set which has 7000 sentences. A 5-gram language model was trained using SRILM toolkit(Stolcke, 2002). The lexicalised reordering models give the orientation of a particular phrase while training. It works as follows. If there is an alignment point found to the left of the current phrase then the orientaton is mono else if the next alignment point is found to the right then the orientation is swap. The models are trained bidirectionally and on both sides. The alignment heuristics used were align-grow-diag-final. More details about GIZA++ can be obtained from (Och and Ney, 2003).

5 Experiments with MOSES parameters

We observed that decreasing the maximum phrase length does not improve the Bleu score. We proceeded to test if the lexicalised reordering models really contributed to the Bleu score. We observed that the removal of the lexicalised reordering models decreased the Bleu score. We also toyed with the distortion limit parameter. We observed that allowing unlimited distortions improves the Bleu Score. Hence the distortion limit was always kept at -1 for allowing unlimited reorderings. We have conducted the following set of experiments. Note that all the experiments which we have conducted can be categorised as pre-processing steps. The motive was to examine if the tapping of all the available rich resources can help improve the Bleu score. Apart from the experiments with parser and bilingual dictionary we also checked whether any other pre-processing steps could improve the Bleu score.

At this stage we have gone for another factored model experiment where, instead of guessing the POS tag through the morph information we go for direct lemma to lemma translation and morph-morph translation. Then the morph information and the lemma are used to guess the POS tag and then the POS tags, morph and the lemma are used to generate the surface forms. Although a signficant improvement has been made still the Bleu score was far away from the baseline. We used the morphological analyzers developed as a part of SHAKTI project(Bharati et al., 2003) to extract the morph information and lemma and POS tags. We also experimented by changing the maximum phrase length to 3 and then 4 as given in(Koehn et al., 2003). But there was no improvement in the Bleu score. We also experimented with the Minimum Bayes Risk Decoder(Kumar et al., 2004) with the Loss Function based on the Bleu scoring function itself.

6 Experiments and Results

We tried out the following set of experiments.

6.1 Learning Transfer Rules from the Source Parsetree

The objective was to incorporate as much as syntactic information to affect the reordering. So we tried to learn the reordering rules from the parallel corpus. We wanted to tap the resources available for English as much as possible. We used Libin's dependency parser(Shen, 2006) to extract the parse tree for English. The obtained parse tree is flattened in this step. The source and the target sentences' POS order are then compared to learn the transfer grammar rules. From the generated rules' set we selected the rules which seemed to be generic. These rules were then applied on the English training data to make the word order as close as possible to Hindi word order. Now the training is done as usual. The following type of rules have been learnt.

```
 VB --- VB
(6)     (6)
 IN   NN-1 --- NN1   IN
 (4)  (5)      (4)    (5)
IN~1_NN&_VB~2 ==> NN&_IN~1_VB~2
```

We achieved the baseline score at this point. We used the additional hindi corpora of 7,000 hindi sentences for training the language model and could get a improvement of 0.65 over the base line. When the dictionary was incorporated in the reordering model, there was a slight improvement of 0.05 in the Bleu score.

6.2 Dictionary Experiments

The English-Hindi bilinugal dictionary was used to translate the unknown words given by the system. The root of the english words was simply replaced by the corresponding Hindi root. When the bilingual dictionary was applied on the unknown words in the output(of the tuned model), there was no improvement in the Bleu score over the baseline model. However an increase in the Bleu score was observed when the dictionary was used to replace the unknown words in the output of the reordering phase. Table 2 shows the results of this experiment without tuning. As expected when the language model was trained using additional corpora there was an increase in the Bleu score.

6.3 Experiments with Phrase Table

We manually checked the translations of the system and found that the translations were interspersed with end-of sentence markers. This means that there was a problem with the phrase extraction heuristic and the phrase translations. We didnot change the heuristic as this was the best heuristic. We instead shifted our focus on the phrase tables. We observed that the invalid translations reduce the Bleu Score. We have included some of these translations below.

System	Without Tuning	With Tuning
Baseline	17.70	20.18
Parser	17.70	-
Dictionary	17.75	-
Phrase Removal	**18.09**	**22.60**

Table 2: Results on EILMT data set

Size of LM → **System ↓**	7500 sentences	8500 sentences
Baseline + Tuning	20.68	20.58
Phrase Removal + Tuning	**21.82**	**20.68**
Parser + Dictionary	18.35	18.60

Table 3: Effect of Additional Language Models

```
 . . . ||| . . . . .
. . . ||| . . . . jo
. . . ||| . . . .
. . . ||| . . . vaha
. . . ||| . . .
. . . ||| .
. . . ||| gaI . . . . jo
```

These translations consisted of the end-of-sentence markers on the source side translated to words on the other side. These type of phrase translations made up 2.5% of the phrase table. To remove these translations from the phrase table, two methods have been proposed. 1) Remove the EOS markers and store them in a hashtable. Train and tune the baseline system. Test the system on the test set and add the EOS markers to the obtained translations. Finally, evaluate the translations against the reference set. 2) Train the system and then remove the invalid translations. Deciding whether a translation is valid or invalid is not a simple yes/no question. This is a debatable issue which leads to the fundamental question of how good is our evaluation stratergy? To avoid this circularity, we have chosen the first procedure. Although the first procedure seems naive, it still outperforms the other techniques as seen from Table 2. We observed that there was a reduction of 10,000 phrases in the size of the phrase table after training.

6.4 Tuning

We tuned our trained model parameters on the development set and then tested the model on the previously used test set. The tuning was done using the Minimum Error Rate Training(Och, 2003) provided in MOSES tool kit itself. We have also conducted experiments using additional language models. Table 3 shows the results of the experiments which involved the language models. We used tourism corpus of size 7500 and 8600 for training language models. No conclusion could be drawn from the results. There is a huge variation in the results due to the change in the language models. We could not figure out any pattern in the results of the language models.

7 Observations

In the course of our experiments, the following observations were made. The Bleu score did not show any significant improvement when the parser was used to reorder the sentences on the source side. This falsifies our initial hypothesis that reordering relations learnt from the parse tree would be more generic and hence would improve the performance. No clear winner emerged from the results. Using a bilingual dictionary did not bring about any change in the Bleu score. An improvement of 27.68% in the Bleu score was observed after the phrase removal experiment. This was the best improvement which we could obtain on the EILMT dataset.

8 Future Directions

Based on the above initial experiments we believe that the reordering of the target language phrases improves substantially by tapping the available resources for English. These experiments are only a first step in improving the Bleu Score and much more has to be achieved.

References

A. Bharati, R. Moona, P. Reddy, B. Sankar, D.M. Sharma, and R. Sangal. 2003. Machine Translation: The Shakti Approach. In *Pre-Conference Tutorial at ICON-2003: International Conference on Natural Language Processing is to be conducted on 19th December*.

P. Koehn and H. Hoang. 2007. Factored translation models. In *Proc. of the 2007 Conference on Empirical Methods in Natural Language Processing (EMNLP/Co-NLL)*.

P. Koehn, F.J. Och, and D. 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 Morristown, NJ, USA.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, volume 45, page 2.

S. Kumar, W. Byrne, JOHNS HOPKINS UNIV BALTIMORE MD CENTER FOR LANGUAGE, and SPEECH PROCESSING (CLSP. 2004. Minimum Bayes-Risk Decoding for Statistical Machine Translation.

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

F.J. 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 Morristown, NJ, USA.

K. Papineni, S. Roukos, T. Ward, and WJ Zhu. 2001. BLEU: a method for automatic evaluation of MT. *Research Report, Computer Science RC22176 (W0109-022), IBM Research Division, TJ Watson Research Center*, 17.

L. Shen. 2006. *STATISTICAL LTAG PARSING*. Ph.D. thesis, University of Pennsylvania.

A. Stolcke. 2002. Srilm-an extensible language modeling toolkit, international conference spoken language processing. *SRI, Denver, Colorado, Tech. Rep*.

S. Venkatapathy and S. Bangalore. Three models for discriminative machine translation using Global Lexical Selection and Sentence Reconstruction. In *SSST*.