Impact of Linguistically Motivated Shallow Phrases in PB-SMT

Santanu Pal[1], Mahammed Hasanuzzaman[2], Sudip Kumar Naskar[1], Sivaji Bandyopadhyay[1]
[1]Department of Computer Science & Engineering, Jadavpur University, Kolkata, India
[2]GREYC, University of Caen Basse-Normandie, France
santanu.pal.ju@gmail.com, mohammed.hasanuzzaman@unicaen.fr, sudip.naskar@cse.jdvu.ac.in, sivaji_cse_ju@yahoo.com,

Abstract

Phrase-based statistical machine translation (PB-SMT) provides the state-of-the-art in machine translation (MT) today. However, unlike syntax-augmented MT systems, it has proven difficult to integrate syntactic knowledge in order to improve translation quality in PB-SMT. This paper describes the effects of linguistically motivated shallow phrases (chunks) incorporated into the state-of-the-art PB-SMT framework for an English−Bangla machine translation task. Linguistically guided phrase pairs are extracted from the training corpus. Afterwards, these phrase pairs are added to the translation model of a PB-SMT system and the probabilities are normalized. We observed that inclusion of these linguistically motivated phrase pairs into the translation model leads to significant improvements in translation quality (**3.18** BLEU points, 29.7% relative) over the baseline system.

1 Introduction

Almost all research in MT being carried out today is corpus based. Statistical techniques using the noisy channel model (Brown et al. 1993) dominate the field and outperform classical ones; however the problem with statistical methods is that they do not employ enough linguistic knowledge to produce a grammatically coherent output (Och et al. 2003). This is because these methods incorporate little or no explicit syntactic knowledge and only captures elements of syntax implicitly via the use of an n-gram language model in the noisy channel framework, which again cannot model long distance dependencies. There has been a long tradition of using syntactic knowledge in statistical machine translation (Wu and Wong, 1998; Yamada and Knight, 2001). After the emergence of phrase-based statistical machine translation (Och and Ney, 2004), several attempts have been made to further augment these techniques with information about the structure of the language. Hierarchical phrase-based modeling (Chiang, 2007) emphasizes the recursive structure of language without concerning itself with the linguistic details. On the other hand, syntax-based modeling uses syntactic categories in addition to recursion, in mapping from source to the target (Galley et al.2004; Zollmann and Venugopal, 2006). While relative improvements over phrase-based baselines have been reported for some language pairs, (Koehn et al., 2003) reported that adding syntax harmed the quality of their SMT system.

From a cognitive point of view phrases are always translated as a whole, while phrases in PBSMT are just a collection of consecutive words, and thus PB-SMT phrases are often not linguistic phrases and do not respect linguistic phrase boundaries. The state-of-art PB-SMT system generally captures phrases implicitly via the use of alignment table which are just n-gram phrases not linguistic phrases.

Traditional PB-SMT system derives phrase pairs directly from the training corpus according to purely statistical method. Thus PB-SMT phrase pair may not follow the syntactic constituents of a sentence; they are just n-grams. This is one of the reasons why PB-SMT often produces ungrammatical translations. Our approach is to restrict the phrase extraction module to extract phrase pairs that begin and end at chunk boundaries. This ensures that the phrase pairs thus extracted thus are not just n-grams; they are linguistically motivated and incorporate syntactic knowledge to some extent.

The remainder of the paper is organized as follows. Next section briefly elaborates the related

work. The English-Bangla PB-SMT system is described in Section 3. Section 4 states the tools and resources used for the various experiments. Section 5 includes the results obtained, together with some analysis. Section 6 concludes and provides avenues for further work.

2 Related Works

Recently, researchers started to investigate how to incorporate syntactic knowledge in PBSMT system to improve the translation quality. (Chiang, 2005) introduced an approach for incorporating syntax into PBSMT, targeting mainly phrase reordering. This was the first work to demonstrate any improvement when adding hierarchical structure to PBSMT. In this approach, hierarchical phrase transduction probabilities are used to handle a range of reordering phenomena in the correct fashion.

(Marcu et al., 2006) presented a similar extension of PBSMT systems with syntactic structure on the target language side. (Zollmann & Venugopal 2006) extended the work introduced in (Chiang, 2005) by augmenting the hierarchical phrases with syntactic categories derived from parsing the target side of the parallel corpus. They associate a target parse tree for each training sentence pair with a search lattice constructed from the existing phrase translations on the corresponding source sentence. Similar to (Chiang, 2005), a chart-based parser with a limited language model has been used.

The problem of SMT system is that phrase pairs are directly extracted form word alignment table; they do not respect linguistic phrases. The extracted phrases of the state-of-art PB-SMT system may contains some words of one phrase and some words of another phrase of the sentences. They do not respect the phrase boundary. But our extraction method is different from the state of the art systems, the extracted phases are either complete linguistic phrases or a combination of linguistic phrases with some specific phrase length. Although Koehn et al. (2003) reported that adding syntactic phrases does not help in improving translation quality we found significant improvements in our case.

3 System Description

3.1 Phrase Extraction

The source language sentences are POS-tagged and chunked. From the chunk labeled sentences, we extract phrases. We modified the Moses phrase extraction scripts of moses so that it extracts phrase pairs which begin and end at chunk boundaries. The phrase pairs extracted thus are made up of chunks.

3.1.1 Source Side Extraction:

The source sentences are chunked using a statistical chunker by considering their POS tags. After chunking the source (i.e., English) sentences, we modify the prepositional chunks so that they include the following noun chunk or a series of noun chunks separated by conjunction. In the below mentioned example the two consecutive chunks "PP (in)" and "NP (1855)" are merged together to form a single PP chunk "PP (in 1855)". The phrase extraction procedure is supplied with maximum allowable phrase length specified by the user. In the following example the maximum phrase length is set to 5. The phrase extraction procedure in the below mentioned example is Overlapping type i.e. a sliding window type where the window slides over the sentence and extracts phrase pairs which begin and end at chunk boundaries following Algorithm 2. We also extract strict n-gram (n=7) linguistic phrase which follows non-sliding type of extraction (cf. Algorithm 1). To avoid too many out of vocabulary phrases we also include all the individual chunks after running Algorithm 1 (Non-overlapping phrase Extraction) and algorithm 2 (Overlapping phrase Extraction).

Source sentence: The republic of Colombia was formally established in 1855.

Source Chunking: NP (The republic) PP (of Colombia) VP (was formally established) PP (in1855) O (.)

Source phrase extraction following Non-overlapping type:
The republic of Colombia
was formally established
In 1855 .

The algorithm 1 takes chunk annotated sentence and maximum phrase length as input. The *for loop* delivers current $chunk_i$ and associated with next $chunk_{i+1}$ ($chunk_j$)and then calculates the length of the $chunk_i$ and $chunk_{i+1}$ (i.e. $chunk_j$) and store in out_phrase and update the i^{th} index for next iteration accordingly, if the calculated length is equals with maximum phrase length. If the calculated length is less than maximum phrase length then next chunk will be concate-

nated with out_phrase until maximum phrase length equals. Otherwise proceed to the next iteration.

ALGORITHM 1: Non-overlaping n-gram Phrase Extraction

```
maxPhL← Max-Phrase-Length;
for i = 1 to m chunks
    out_phrase ← chunk_i;
    length ← number words in chunk_i;
    j = i+1
    C ← chunk_j;
    P ← set of words in C;
    L← number words in P;
    if (length + L)<=maxPhL
                Concatenate C with out_phrase;
                length← length +L;
     else
          add out_phrase into List;
          out_phrase←null
     end if
end for
```

The Algorithm (Overlapping) 2 is slightly different from Algorithm 1 (Non-overlapping); it takes chunk annotated sentence and maximum phrase length as input. The *outer for loop* delivers current $chunk_i$ to the *inner for loop*. The *inner for loop* reads next $chunk_{i+1}$ and then calculates the length of the $chunk_i$ and $chunk_{i+1}$ (i.e. $chunk_j$) and store in out_phrase, if the calculated length is equals with maximum phrase length. If the calculated length is less than maximum phrase length then $chunk_{j+1}$ concatenated with out_phrase until maximum phrase length equals. Otherwise, the i^{th} index will be incremented to finding next overlapped chunk. In this extraction also, we set a value maximum phrase length, and for each chunk we try to merge as many chunks as possible so that the number of tokens in the merged chunk never exceeds maximum phrase length.

Source phrase extraction following Overlapping type:
The republic of Colombia
Of Colombia was formally establish
Was formally establish in 1855
In 1855 .

ALGORITHM 2: Overlapping n-gram Phrase Extraction

```
maxPhL← Max-Phrase-Length;
for i = 1 to m chunks
    out_phrase ← chunk_i;
    length ← number of words in chunk_i;
    for j = i+1 to m chunks
          C ← chunk_j;
          P ← set of words in C;
          l← number words in P;
          if (length + l)<=maxPhL
                Concatenate C with out_phrase;
                length← length + l;
          end if
          add out_phrase into List;
    end for
    out_phrase←null
end for
```

Target sentence chunking:

NP(কলম্বিয়ার প্রজাতন্ত্র) NP(1855 খ্রীষ্টাব্দে) RBP(আনুষ্ঠানিকভাবে) JJP(প্রতিষ্ঠিত) VP(হয়েছিল) (.)

Source target word Alignment:
0-0 3-0 1-1 2-1 5-2 8-3 7-4 8-4 6-5 6-6 9-7

Source Sentence:

[The(0) republic(1)] [of(2) Colombia(3)] [was(4) formally(5) established(6)] [in(7) 1855(8)] [.(9)]

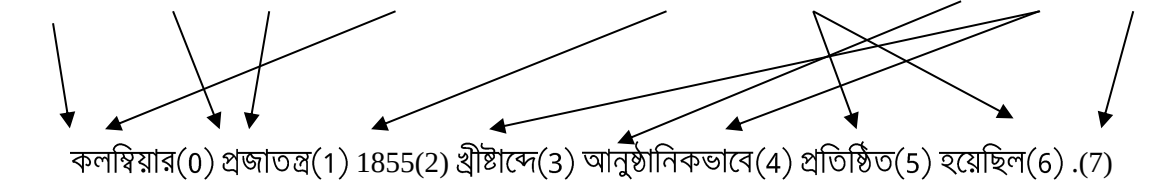

Figure 1 Word alignment provided by GIZA++

3.1.2 Source-target Phrase Extraction files creation:

Using the alignment file provided by GIZA++, we created an alignment file using grow-diag-final-and algorithm and created two files as directed in moses (Koehn, 2009) toolkit - extract.direct and extract.inv. The below mentioned example shows, the steps of phrase alignment file creation procedure by using the knowledge of word alignment produced by GIZA++. Using these two extracted files we have created phrase table following a similar method described in Koehn, 2003. The above extracted phrase alignment file are pruned and discarded those phrases that contains extra words on either source or target phrase contains extra word that is not relevant to the phrase alignment.

The above examples the phrase alignment no 6 is marked as erroneous because "*the republic*" has already aligned with "প্রজাতন্ত্র (prajatantra)" but in left hand side of phrase 6 does not contains "*the republic*".

Phrase Alignment file

1. The republic ||| প্রজাতন্ত্র
2. Of Columbia ||| কলম্বিয়ার
3. Was formally established ||| প্রতিষ্ঠিত হয়েছিল
4. In 1855 ||| 1855 খ্রীষ্টাব্দে আনুষ্ঠানিকভাবে
5. The republic of Colombia ||| কলম্বিয়ার প্রজাতন্ত্র
6. Of Colombia was formally established ||| কলম্বিয়ার প্রজাতন্ত্র 1855 খ্রীষ্টাব্দে আনুষ্ঠানিকভাবে প্রতিষ্ঠিত হয়েছিল
7. Was formally establish in 1855 ||| 1855 খ্রীষ্টাব্দে আনুষ্ঠানিকভাবে প্রতিষ্ঠিত হয়েছিল
8. In 1855 . ||| 1855 খ্রীষ্টাব্দে আনুষ্ঠানিকভাবে .

Example 1 Phrase Alignment file

3.2 Phrase table:

We constrained the PBSMT phrase length to a maximum of 7 and a minimum of 4. After extracting all phrase pairs we calculated lexical weighting and phrase translation probabilities in both source and target direction.

The phrase translation probability corresponding to the phrase pair (f, e) is given by equation [1].

$$\emptyset(f|e) = \frac{count(f,e)}{\sum_{f_i}(f,e)} \quad [1]$$

Lexical weighting (*Lw*) is given by equation [2].

$$Lw(f|e,a) = \prod_{i=1}^{length(f)} \frac{1}{\{j|(i,j) \in a\}} \sum_{\forall(i,j)\in a} w(fi|ej) \quad [2]$$

where $w(f_i|e_j)$ signifies word translation probability.

To speed up decoding, we integrated phrases associated with their phrase translation probability and lexical weighting into the phrase table.

3.3 Decoder:

We have used the state-of-the-art moses decoder to decode the test sentences. The moses decoder is initialized with an empty hypothesis; a new hypothesis is expanded by a sequence of untranslated foreign words and a possible target phrase translation is selected. The decoder essentially extracts translations from the translation table and recombines all the fragment translations and updates the hypothesis. At the end it produces hypotheses ranked according to their probability mass considering all the models.

4 Tools and Resources

We carried out our experiments with a sentence-aligned English−Bangla parallel corpus containing 23,492 parallel sentences from the travel and tourism domain. The corpus was collected from the EILMT project. The the CRF chunker[1] together with Stanford Parser[2] was used for identifying individual chunks in the source side of the parallel corpus.

The sentences on the target side (Bangla) are parsed and chunks are extracted using tools obtained from the IL-ILMT[3] project. The effectiveness of the linguistically motivated phrase table is demonstrated by using the standard log-linear PB-SMT model, GIZA++ implementation of IBM word alignment model 4, phrase-extraction heuristics described in (Koehn et al., 2003), minimum-error-rate training (Och, 2003) on a held-out development set, language model trained using SRILM toolkit (Stolcke, 2002) with interpolated modified Kneser-Ney smoothing (Kneser and Ney, 1995) and the Moses decoder (Koehn et al., 2007).

[1] http://crfchunker.sourceforge.net/

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

[3] The EILMT and ILILMT projects are funded by the Department of Information Technology (DIT), Ministry of Communications and Information Technology (MCIT), Government of India.

5 Experiment and Result

We randomly identified 500 sentences each for the development set and the test set from the initial parallel corpus. The rest were considered as the training corpus. The training corpus was filtered with the maximum allowable sentence length of 100 words and sentence length ratio of 1:2 (either way). The filtered training corpus contains 22,492 sentences. In addition to the target side of the parallel corpus, a monolingual Bangla corpus containing 488,086 words from the tourism domain was used for the target language modeling.

5.1 Baseline System setup

Baseline experiments were carried out with different n-gram settings for the language model and the maximum phrase length, from which we found that a 5-gram language model and a maximum phrase length of 7 produce the optimum baseline result. The rest of the experiments were carried out using these settings.

5.2 Experiments

The experiment has been carried out in two directions: (i) Linguistically Motivated PB-SMT has been evaluated in various phrase length setting which has been reported in table 1, experiment number 2 and 3, (ii) We have incorporated linguistically motivated extracted phrases with the extracted phrases given by state-of-art baseline PB-SMT system and normalized their probabilities. Our experiments show that high level of performance is achieved with fairly simple change in the phrase extraction procedure in terms of incorporating linguistically motivated phases. In Experiment no. 2, the score has decreased to 0.6 BLUE matric when we have considered minimum phrase length (PL) 1, but when we setup minimum phrase length 4, the score has been increased **1.12** BLEU matric point. In the second direction of experiment (4, 5, 6, 7), we extracted phrases in strict (non-Overlap basis) and integrating with baseline PB-SMT system. We got significant improvement over baseline, but got maximum improvement only at PL 4 with strict basis. But when we set up our experiment by integrating Overlap type (sliding window) of extraction with the baseline system, we achieved **3.18** BLUE point (**29.7%**) relative improvements over baseline system, because linguistically motivated phrases are gaining more weight in compare with state-of-art baseline PB-SMT.

Thus, extrinsic evaluation was carried out on the MT quality using the well-known automatic MT evaluation metrics: BLEU (Papineni et al., 2002) and NIST (Doddington, 2002). Bangla is a morphologically rich language and has relatively free phrase order. Proper evaluation of the English-Bangla MT evaluation ideally requires multiple set of reference translations. Moreover, the training set was smaller in size.

Exp. No.	Experiments	BLEU	NIST
1	Baseline PB-SMT	10.68	4.12
2	Linguistically Motivated(LM) -Overlap PB-SMT[1-7]	10.31	4.08
3	LM-Overlap-PB-SMT[4-7]	11.80	4.24
4	LM-non-Overlap-PB-SMT[PL-4 strict]+baseline PBSMT	**12.75***	**4.38**
5	LM-non-Overlap-PBSMT[PL-5 strict]+baseline PBSMT	12.35	4.38
6	LM-non-Overlap-PBSMT[PL-6 strict]+baseline PBSMT	11.58	4.36
7	LM-non-Overlap-PBSMT[PL-7 strict]+baseline PBSMT	11.53	4.36
8	LM-Overlap-PBSMT[PL-4-7]+baseline PBSMT	**13.86**†	**4.41**

Table 1: Evaluation result obtained by integrating linguistically motivated phrases in PB-SMT system, * Significant improvement using non-overlapped PB-SMT System o, † Significant improvement using overlapped PB-SMT System over baseline PB-SMT

6 Conclusion and Future Work

In this paper we presented a system which shows how linguistically motivated shallow phrases can improve the performance of PB-SMT system on an English−Bangla translation task. Our best system yields 3.18 BLEU points improvement over the baseline, a 29.7% relative increase. We compared a subset of the output of our best system with that of the baseline system, and the output of our best system in most cases looked better in terms of grammaticality. Further work will be carried out by constraining the moses decoder to consider only linguistically motivated phrases.

References

Andreas Zollmann and Ashish Venugopal. 2006. Syntax augmented machine translation via chart parsing. *In Proceedings of WMT 2006*.

D. Marcu, W. Wang, A. Echihabi, and K. Knight. 2006. SPMT: Statistical Machine Translation with Syntactified Target Language Phrases. *In Proceedings of EMNLP-2006*, Sydney, Australia, pp.44–52, 2006.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. *In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL-2005)*.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2).

Dekai Wu, and Hongsing Wong. 1998. Machine translation with a stochastic grammatical channel. *In Proceedings of ACL -1998*.

Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)* 39 (1): pp. 1–38.

Doddington, George. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. *In Proc. of the Second International Conference on Human Language Technology Research (HLT-2002)*, San Diego, CA, pp. 128-132.

Franz Och, and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4).

Hany Hassan. 2006. Syntactic phrase-based statistical machine translation. *Spoken Language Technology Workshop, 2006*. IEEE. IEEE.

Kenji Yamada, and Kevin Knight. 2001. A syntax-based statistical translation model. *In Proceedings of ACL-2001*.

Kneser, Reinhard, and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. *In Proc. of the IEEE Internation Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, Detroit, MI, pp. 181–184.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? *In Proceedings of HLT-NAACL*.

Och, Franz J. 2003. Minimum error rate training in statistical machine translation. *In Proc. of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-2003)*, Sapporo, Japan, pp. 160-167.

Och, Franz Josef, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev. 2003. Syntax for statistical machine translation. *Final Report of Johns Hopkins 2003 Summer Workshop*.

P.E. Brown, S.A.D. Pietra, V.J.D. Pietra, R.L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, Vol. 16.N. 2, pp. 79-85.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. *In Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002)*, Philadelphia, PA, pp. 311-318

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. *In Proc. of HLT-NAACL 2003: conference combining Human Language Technology conference series and the North American Chapter of the Association for Computational Linguistics conference series*, Edmonton, Canada, pp. 48-54.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. *In Proc. of the 45th Annual meeting of the Association for Computational Linguistics (ACL 2007): Proc. of demo and poster sessions*, Prague, Czech Republic, pp. 177-180.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In EMNLP-2004: *Proc. of the 2004 Conference on Empirical Methods in Natural Language Processing*, 25-26 July 2004, Barcelona, Spain, pp 388-3.

Vogel, Stephan, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. *In Proc. of the 16th International Conference on Computational Linguistics (COLING 1996)*, Copenhagen, pp. 836-841.

Xuan-Hieu Phan. 2006. *CRFChunker: CRF English Phrase Chunker*. http://crfchunker.sourceforge.net/.

Zens, Richard, Franz Josef Och, and Hermann Ney. Phrase-based statistical machine translation. 2002. KI 2002: *Advances in Artificial Intelligence*. Springer Berlin Heidelberg, 2002. 18-32.