Handling Multiword Expressions in Phrase-Based Statistical Machine Translation

Santanu Pal, Tanmoy Chakraborty , Sivaji Bandyopadhyay
Department of Computer Science & Engineering
Jadavpur University
santanu.pal.ju@gmail.com, its_tanmoy@yahoo.co.in,
sivaji_cse_ju@yahoo.com

Abstract

Preprocessing of the parallel corpus plays an important role in improving the performance of a phrase-based statistical machine translation (PB-SMT). In this paper, we propose a frame work in which predefined information of Multiword Expressions (MWEs) can boost the performance of PB-SMT. We preprocess the parallel corpus to identify Noun-noun MWEs, reduplicated phrases, complex predicates and phrasal prepositions. Single-tokenization of Noun-noun MWEs, phrasal preposition (source side only) and reduplicated phrases (target side only) provide significant gains over our previous best PB-SMT model. Automatic alignment of complex predicates substantially improves the overall MT performance and the word alignment quality as well. For establishing NE alignments, we transliterate source NEs into the target language and then compare them with the target NEs. Target language NEs are first converted into a canonical form before the comparison takes place. The proposed system achieves significant improvements (6.38 BLEU points absolute, 73% relative improvement) over the baseline system on an English- Bengali translation task.

1 Introduction

Performance of a Statistical machine translation (SMT) system depends mainly upon the good quality word and phrase alignment tables that constitute the translation knowledge acquired from a parallel corpus. In this paper, we show that handling the Multiword Expressions (MWE) can improve the performance of a SMT system.

The structure and meaning of MWEs cannot be derived from their component words, as they occur independently. Examples include conjunctions (*'as well as'*), idioms (*'kick the bucket'* means 'to die'), phrasal verbs (*'find out'*), compound noun (*'building complex'*), phrasal preposition ('a*ccording to'*) etc. Briefly, MWE can be roughly defined as idiosyncratic interpretations that cross word boundaries (Sag et al., 2002). Complex Predicates (*CPs*) are made of *verb + verb* (*compound verbs*) or *noun/adjective/adverb* +verb (*conjunct verbs*) patterns in most of the *South Asian language*s like Bengali, Hindi etc. As Bengali is morphologically enriched, the morphological knowledge is required to identify Complex Predicates (*CPs*). The following are some example Complex Predicates (*CPs*) in Bengali: compound verbs (e.g., মেরে ফেলা *mere phela* 'kill', বলতে লাগল *bolte laglo* 'started saying') and conjunct verbs (e.g., ভরসা করা *bharsha kara* 'to depend', ঝকঝক করা *jhakjhak kara* 'to glow').

The first verb in a compound verb is called as *Full Verb* which is represented either as conjunctive participial form -এ *–e* or the infinitive form -তে *–te* at the surface level. The other verb bears the inflection based on *Tense*, *Aspect* and *Person*. These *Light Verbs* (*LV*) are polysemous, semantically bleached and confined into some definite candidate seeds (Paul, 2010).

On the other hand, each Bengali conjunct verb (*ConjV*) consists of noun, adjective or adverb followed by a *Light Verb* (*LV*). The *Light Verbs*

(*LVs*) bear the appropriate inflections based on *Tense*, *Aspect* and *Person*.

Since the conventional meaning of the *Light Verbs* in Complex Predicates (*CPs*) is absent, such complex predicates are considered as Multi Word Expressions (*MWEs*)(Baldwin and Kim, 2010, Sinha, 2009). The other types of predicates such as নিয়ে গেল *niye gelo* 'take-go' (took and went), দিয়ে গেল *diye gelo* 'give-go' (gave and went) follow the lexical pattern *FV+LV*, similar to Complex Predicates (*CPs*) but the *Full Verb* and *Light Verb* behave as independent syntactic entities. These verb patterns are non-Complex Predicates (*non-CPs*) and are also termed as *Serial Verb* (*SV*) (Mukherjee *et al.*, 2006).

Traditional approaches to word alignment follow the IBM Models (Brown et al., 1993). These approaches are unable to handle many-to-many alignments and hence do not work well with multi-word expressions, especially with NEs, reduplications and complex predicates. The alignment probabilities in the well-known Hidden Markov Model (HMM: Vogel et al., 1996) depend on the alignment position of the previous word. The HMM model does not explicitly consider many-to-many alignments.

In this experiment, we address this many-to-many alignment problem indirectly. Our objective is to see how the identification of MWEs enhances the performance of the SMT system. In this work, several types of MWEs like phrasal prepositions and Verb-object combinations are automatically identified on the source side while named-entities and complex predicates are identified on both sides of the parallel corpus. In the target side, identification of the Noun-noun MWEs and reduplicated phrases are carried out. We use simple rule-based and statistical approaches to identify these MWEs. Source and target language NEs are aligned using a statistical transliteration technique. We rely on these automatically aligned NEs and treat them as translation examples (Pal.et.al, 2010). Adding bilingual dictionaries, which in effect are instances of atomic translation pairs, to the parallel corpus is a well-known practice in domain adaptation in SMT (Eck et al., 2004; Wu et al., 2008). We modify the parallel corpus by converting the MWEs into single tokens and adding the aligned NEs and complex predicates in the parallel corpus to improve the word alignment and hence the phrase alignment quality. The preprocessing of the parallel corpus results in improved MT quality in terms of automatic MT evaluation metrics.

The remainder of the paper is organized as follows. Next section briefly elaborates the related work. The English-Bengali PBSMT 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 Work

Moore (2003) used capitalization cues for identifying NEs on the English side and then applied statistical techniques to decide which portion of the target language corresponds to the specified English NE. A Maximum Entropy model based approach for English—Chinese NE alignment has been proposed in Feng et al. (2004) which significantly outperforms IBM Model 4 and HMM. A method for automatically extracting NE translingual equivalences between Chinese and English based on multi-feature cost minimization has been proposed in Huang et al. (2003).

Venkatapathy and Joshi (2006) reported a discriminative approach to use the compositionality information of verb-based multi-word expressions in order to improve the word alignment quality. Ren et al. (2009) presented a log likelihood ratio based hierarchical reducing algorithm to automatically extract bilingual MWEs. They investigated the usefulness of these bilingual MWEs in SMT by integrating bilingual MWEs into the Moses decoder (Koehn et al., 2007). They observed the highest improvement with an additional feature that identifies whether or not a bilingual phrase contains bilingual MWEs. This approach was generalized in Carpuat and Diab (2010) who replaced the binary feature by a count feature representing the number of MWEs in the source language phrase.

Intuitively, MWEs on the source and the target sides should be both aligned in the parallel corpus and translated as a whole. However, in the state-of-the-art PB-SMT systems, the constituents of an MWE are marked and aligned as parts of consecutive phrases, since PB-SMT (or any other approaches to SMT) does not generally treat MWEs as special tokens. Another problem with SMT systems is the wrong translation of verb phrases. Sometimes verb phrases are deleted in the output sentence. Moreover, the words inside verb phrases are generally not aligned one-to-one; the alignments of the words inside source and target verb phrases are mostly many-to-many, particularly so for the English—Bengali language pair. These are the motivations behind considering MWEs like NEs, reduplicated phrases, prepositional phrase and compound verbs for special treatment in this work.

By converting the MWEs into single tokens, we make sure that PB-SMT also treats them as a whole. The first objective of the present work is to see how single tokenization and alignment of NEs on both the sides, single tokenization of phrasal verbs and phrasal prepositions on the source side and single tokenization of reduplicated phrases and noun-noun compounds on the target side affects the overall MT quality. The second objective is to see whether prior automatic alignment of complex predicates and single-tokenized MWEs can bring any further improvement in the overall performance of the MT system.

We carried out the experiments on English—Bengali translation task. Bengali shows high morphological richness at lexical level. Language resources in Bengali are not widely available. Furthermore, this is the first time when the identification of MWEs in Bengali language is used to enhance the performance of an English-Bengali Machine Translation System.

3 System Description

3.1 PB-SMT

SMT models machine Translation as a decision process. The translation $e_1^I = e_1 \ldots e_i \ldots e_I$ of a source sentence $f_1^J = f_1 \ldots f_j \ldots f_J$ is chosen to maximize the following equation (1):

$$\arg\max_{I,e_1^I} P(e_1^I \mid f_1^J) = \arg\max_{I,e_1^I} P(f_1^J \mid e_1^I).P(e_1^I) \ldots (1)$$

where $P(f_1^J \mid e_1^I)$ and $P(e_1^I)$ denote the translation model and the target language model (Brown et al., 1993) respectively. In log-linear phrase-based SMT, the posterior probability $P(e_1^I \mid f_1^J)$ is directly modeled as a log-linear combination of features (Och and Ney, 2002), that usually comprise M translational features, and the language model, as in equation (2):

$$\log P(e_1^I \mid f_1^J) = \sum_{m=1}^{M} \lambda_m h_m(f_1^J, e_1^I, s_1^K) + \lambda_{LM} \log P(e_1^I) \ldots\ldots. (2)$$

where $s_1^k = s_1 \ldots s_k$ denotes a segmentation of the source and target sentences respectively into the sequences of phrases $(\hat{e}_1, \ldots, \hat{e}_k)$ and $(\hat{f}_1, \ldots, \hat{f}_k)$ as shown (we set $i_0 = 0$) in equation (3).

$$\forall 1 \le k \le K,\ s_k = (i_k, b_k, j_k), \quad \hat{e}_k = e_{i_{k-1}+1} \ldots e_{i_k}, \quad \hat{f}_k = f_{b_k} \ldots f_{j_k} \ldots\ldots\ldots\ldots\ldots\ldots (3)$$

Each feature $\hat{h}_m$ in equation (2) can be rewritten as in equation (4):

$$h_m(f_1^J, e_1^I, s_1^K) = \sum_{k=1}^{K} \hat{h}_m(\hat{f}_k, \hat{e}_k, s_k) \ldots\ldots\ldots.. (4)$$

where $\hat{h}_m$ is a feature that applies to a single phrase-pair. It thus follows:

$$\sum_{m=1}^{M} \lambda_m \sum_{k=1}^{K} \hat{h}_m(\hat{f}_k, \hat{e}_k, s_k) = \sum_{k=1}^{K} \hat{h}(\hat{f}_k, \hat{e}_k, s_k) \ldots (5)$$

where $\hat{h} = \sum_{m=1}^{M} \lambda_m \hat{h}_m$.

3.2 Preprocessing of the parallel corpus

The initial English-Bengali parallel corpus is cleaned and filtered using a semi-automatic process. We employed several kinds of multiword information: Noun-noun MWEs, reduplicated phrases, complex predicates, phrasal prepositions, verb-object combination and NEs. Compound verbs are first identified on both sides of the parallel corpus. Das et al. (2010) analyzed and identified a category of compound verbs (*Verb* + *Verb*) and conjunct verbs (*Noun /Adjective/Adverb* + *Verb*) for Bengali. We adapted their strategy for identification of compound verbs as well as serial verbs (*Verb* + *Verb* + *Verb*) in Bengali.

For the identification of Named-Entities and their alignment, we have adopted a similar technique as reported in Pal.et.al (2010). Reduplicated phrases do not occur very frequently in the English corpus; some of them (like correlatives, semantic reduplications) are not found in English (Chakraborty and Bandyopadhyay, 2010).

But reduplication plays a crucial role on the target Bengali side as they occur with high frequency. These reduplicated phrases are considered as a single-token so that they may map to a single word on the source side. Phrasal prepositions and verb object combinations are also treated as single tokens. Once the compound verbs and the NEs are identified on both sides of the parallel corpus, they are assembled into single tokens. When converting these MWEs into single tokens, we replace the spaces with underscores ('_'). Since there are already some hyphenated words in the corpus, we do not use hyphenation for this purpose. Besides, the use of a special word separator (underscore in our case) facilitates the job of deciding which single-token MWEs to be de-tokenized into its constituent words, before evaluation.

3.3 MWE Identification in Source Side

We have adopted the UCREL[1] Semantic analysis System (USAS) developed by Lancaster University (Rayson et al., 2004). The USAS is a software tool for the automatic semantic analysis of English spoken and written data. Various types of Multi-Word Units (MWU) that are identified by the USAS software include: verb-object combinations (e.g. *stubbed out*), noun phrases (e.g. *riding boots*), proper names (e.g. *United States of America*), true idioms (e.g. *living the life of Riley*) etc. The USAS software has a reported precision value of 91%.

3.4 MWE Identification in Target Side

3.4.1 Noun-Noun MWE Identification

In the past few years, noun compounds have received increasing attention as researchers work towards the goal of full text understanding. Compound nouns are nominal compounds where two or more nouns are combined to form a single phrase such as *'golf club'* or '*computer science department'* (Baldwin and Kim, 2010). Compound noun MWEs can be defined as a lexical unit made up of two or more elements, each of which can function as a lexeme independent of the others(s) in different contexts. Compound noun MWEs show some phonological and/or grammatical isolation from normal syntactic usage. In English, Noun-Noun (NN) compounds occur with high frequency and high lexical and semantic variability (Tanaka and Baldwin, 2003). In this experiment, we have used simple statistical methods for identifying Noun-noun MWEs. The system uses Point-wise Mutual Information (PMI), Log-likelihood Ratio (LLR) and Phi-coefficient, Co-occurrence measurement and Significance function (Agarwal et al. 2004) measures. Final evaluation has been carried out by combining the results of all the methods. A predefined cut-off score has been considered and the candidates having scores above the threshold value have been considered as MWEs.

3.4.2 Identification of Reduplication

In all languages, the repetition of noun, pronoun, adjective and verb are broadly classified under two coarse-grained categories: repetition at the (a) ***expression level*** and at the (b) ***contents or semantic level.*** The repetition at both the levels is mainly used for emphasis, generality, intensity or to show continuation of an act. In this experiment, we have used a simple rule-based approach (Chakraborty and Bandyopadhyay, 2010) to identify reduplication in the Bengali-side corpus. In that approach, the authors have classified expression-level Bengali reduplication into five fine-grained subcategories: (i) Onomatopoeic expressions (**khat khat**, *knock knock*), (ii) Complete Reduplication (**bara-bara,** *big big*), (iii) Partial Reduplication (**thakur-thukur,** *God*), (iv) Semantic Reduplication (**matha-mundu**, *head*) and (v) Correlative Reduplication (**mara-mari**, *fighting*). The present work tries to cover almost all the above mentioned types. We have used simple rules and morphological properties at lexical level. The Bengali monolingual dictionary has been used for identification of semantic reduplications.

3.5 Automatic Alignment of NEs and Complex Predicates

An NE parallel corpus is created by extracting the source and the target (single token) NEs from the NE-tagged parallel corpus and aligning the NEs using the strategies as applied in (Pal.et.al, 2010). Extraction and alignment of complex predicates have been carried out using the procedures discussed in the following subsections.

[1] http://www.comp.lancs.ac.uk/ucrel

3.5.1 Complex predicate Extraction

For the extraction of Complex Predicates (CPs) in the target side corpus, we have focused on compound verbs (CPs) (*Verb* + *Verb*) and conjunct verbs (*Noun /Adjective/Adverb* + *Verb*) and have adopted the method applied in Das et.al, (2010). Serial verbs (*Verb* + *Verb* or *Verb* + *Verb* + *Verb* patterns) have also been considered for extraction in the present work. Individual verbs in a serial verb identify separate actions while those in a compound verb convey a single action together. Verbs that are first identified as part of a serial verb are not considered further while identifying compound verbs. Examples of some complex predicates and serial verbs in Bengali and the associated English source as identified in the present work by the complex predicate identification system are shown below.

দেখা_যায় (*dekha jai*) /SV [can be viewed]
নিয়ে_যেতে_পারেন (*niya_jete_paren*) /SV [can carry]
অবরোধ_করতে_পারত (*aborodh_korte_parto*) / ConjV [would have blocked]
পাড়ি_দেয় (*paRi_dei*) / CompV [arrived]
চেখে_দেখুন (*Chekhe_dekhun*) / CompV [test]

(*SV: = Serial Verb, ConjV: = Conjunct Verb, CompV:=Compound Verb*)

At first, we have extracted all serial verbs and complex predicates with their sentence ids from the target side. We have also extracted the entire verb chunk from the corresponding source side in the aligned pair.

3.5.2 Verb Chunk / Complex Predicate Alignment

A. Initial source and target alignment: Initially, it is assumed that all the members of the English verb chunk in an aligned sentence pair are aligned with the members of the Bengali complex predicates. The following example illustrates the point:

designed ||| তৈরি_হয় (*toiri_hoi)*
designed ||| নকশা_করা (*noksha kora*)
is ||| তৈরি_হয় (*toiri_hoi*)
is ||| নকশা_করা (*noksha kora*)
was built ||| তৈরি_হয় (*toiri_hoi*)
was built ||| নকশা_করা(*noksha kora*)

B. Statistical Aligner: The verb chunk and the complex predicate alignments in the initial alignment list, in which only one verb chunk is present in an English side sentence, are considered as correct alignments. This process is carried out in an iterative manner in which the correct alignments identified in first iteration are used to locate further correct alignments in the next iteration. For all other alignments, synonyms of the English verb chunks are identified in the English WordNet 3.0. Each such synonym verb is aligned with the same Bengali complex predicate generating a number of additional alignments. Some of these additional alignments might have already occurred in the alignment list. Hence, frequencies of the various alignments after the expansion using the synonym verbs are counted. The root forms of the English and the Bengali verbs have been considered in this process. Those alignments whose frequencies fall below some heuristically set cut-off mark are put in a separate doubtful alignment list.

Eg:
adorned ||| অলংকৃত_করেছিল/ConjV ||| 0.26
allowing ||| অনুমতি_দেয়/ConjV ||| 0.2308
allowing ||| অনুমতি_দিয়েছিলেন/ConjV ||| 0.2308
test ||| চেখে_দেখুন/CompVerb ||| 0.5333

The above example specifies the output of the statistical aligner. The last column signifies the probability of the corresponding alignment frequency.

C. Pattern generator and Aligner: The pattern generator extracts patterns for both the source and the target side tokens from the generated correct alignment list. The root form of the main verb, auxiliary verb present in the verb chunk and the associated tense, aspect and modality information are extracted for the source side token. Similarly, root form of the Bengali verb and the associated vibhakti (inflection) are identified on the target side token. Similar patterns are extracted for each alignment in the doubtful alignment list. This list has been generated by eliminating the alignment list that was provided by the statistical aligner from the initial source-target alignment list.

Each pattern alignment for the entries in the doubtful alignment list is checked with the patterns identified in the correct alignment list. If both the source and the target patterns for a doubtful alignment match with the source and the target patterns of a correct alignment, then the doubtful alignment is considered as a correct one.

The doubtful alignment list is checked again to look for a single doubtful alignment for a sentence pair. Such doubtful alignments are considered as correct alignment.

4 Tools and Resources used

A sentence-aligned English-Bengali parallel corpus containing 14,187 parallel sentences from the travel and tourism domain has been used in the present work. The corpus has been collected from the consortium-mode project “Development of English to Indian Languages Machine Translation (EILMT) System[2]”. The Stanford Parser[3], Stanford NER, CRF chunker[4] and the Wordnet 3.0[5] have been used for identifying complex predicates in the source English side of the parallel corpus.

The sentences on the target side (Bengali) are POS-tagged by using the tools obtained from the consortium mode project “Development of Indian Language to Indian Language Machine Translation (IL-ILMT) System[2]”. NEs in Bengali are identified using the NER system of Ekbal and Bandyopadhyay (2008). We have used the Stanford Parser and the Bengali NER.

The effectiveness of the MWE-aligned parallel corpus is demonstrated by using the standard log-linear PB-SMT model as our baseline system: 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, target language model trained using SRILM toolkit (Stolcke, 2002) with Kneser-Ney smoothing (Kneser and Ney, 1995) and the Moses decoder (Koehn et al., 2007).

5 Experiments and Results

We have randomly identified 500 sentences each for the development set and the test set from the initial parallel corpus. The rest are 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). Finally the training corpus contained 13,176 sentences. In addition to the target side of the parallel corpus, a monolingual Bengali corpus containing 293,207 words from the tourism domain was used for the target language model. We experimented with different n-gram settings for the language model and the maximum phrase length and found that a 4-gram language model and a maximum phrase length of 4 produce the optimum baseline result. We carried out the rest of the experiments using these settings.

Training set	English		Bengali	
	T	U	T	U
CPs	4874	2289	14174	7154
reduplicated word	-	-	85	50
Noun-noun compound	892	711	489	300
Phrasal preposition	982	779	-	-
Phrasal verb	549	532	-	-
Total NE words	22931	8273	17107	9106

Table 1. MWE Statistics. (T - Total occurrence, U – Unique, CP – complex predicates)

The system continues with the various preprocessing of the corpus. Our hypothesis is that as more and more MWEs are identified and aligned properly, the system shows the improvement in the translation procedure. Table 1 shows the MWE statistics of the parallel training corpus. It is observed from Table 1 that NEs occur with high frequency in both sides compared to other types of MWEs. It suggests that prior alignment of the NEs plays a role in improving the system performance.

Single tokenization of NEs of any length on both the sides followed by GIZA++ alignment has given a huge impetus to system performance

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

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

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

[5] http://wordnet.princeton.edu/

(4.59 BLEU points absolute, 52.5% relative improvement over the baseline). In the source side, the system treats the phrasal prepositions, verb-object combinations and noun-noun compounds as a single token. The best performance of the system (13.99 BLEU score) with source side treatment and named entity alignment (NEA) is achieved with phrasal prepositions and verb-object combinations as single tokens. But single tokenization of the noun-noun compounds and incorporating them in the system, degrades the performance. The accuracy of the UCREL semantic toolkit used for extracting the English noun-noun compounds is not satisfactory especially for the tourism domain. Noun-noun compounds exhibit a many-to-many alignment between the source and the target sides. Sometimes, English noun-noun compounds may not be translated as noun-noun compounds in Bengali.

Experiments		**Exp**	**BLEU**	**NIST**
Baseline		1	**8.74**	**3.98**
Baseline Best System (Alignment of NEs of any length (NEA))		2	**13.33**	**4.44**
Source Side Treatment + NEA	Phrasal preposition as single-token (SPPaST)	3	13.76	4.39
	Verb-object combination as a single-token (SVOaST)	4	13.61	4.40
	Verb-object combination and phrasal preposition as a single-token (SPPaST+SVOaST)	**5**	**13.99**	**4.41**
	Noun-noun compound as Single token (SNNaST)	6	13.61	4.40
	(SPPaST+SNNaST)	7	13.71	4.41
	(SPPaST+SNNaST+SVOaST)	8	13.89	4.42
Target Side Treatment+ NEA	Reduplicated phrase as single-token (TRWaST)	**9**	**13.84**	**4.42**
	Noun-noun compound as Single token (TNNaST)	10	13.75	4.42
	Reduplicated word and Noun noun compound as single-token (TRWaST + TNNaST)	11	13.83	4.42
Both Side Treatment+ NEA	SPPaST+TRWaST	12	14.07	4.41
	SPPaST+TRWaST+TNNaST	13	14.38	4.43
	SPPaST+SNNaST+TRWaST+TNNaST	14	14.20	4.43
	SPPaST+SVOaST+TRWaST+TNNaST	**15**	**14.58**	**4.44**
	SPPaST+SNNaST+SVOaST+TRWaST+TNNaST	16	14.51	4.43
Baseline Best System	Complex predicates alignment (CPA)	17	14.14	4.43
Baseline Best System	CPA+ (Best combination)SPPaST+SVOaST+TRWaST+TNNaST†	**18**	**15.12**	**4.48**

Table 2. Evaluation results for different experimental setups. (The '†' marked systems produce statistically significant improvements on BLEU over the baseline system)

In the target side, single tokenization of reduplicated phrases and noun-noun compounds has been done followed by alignments using the GIZA++ tool. The best performance of the system (13.84 BLEU score) with target side treatment and named entity alignment (NEA) is achieved with reduplicated phrases as single tokens. A BLEU score of 13.83 has been obtained when both reduplicated phrases and noun-noun-compounds have been identified as single tokens.

The system achieves the best performance (14.58 BLEU score) when phrasal prepositions and verb-object combinations are single tokenized on the source side, reduplicated phrases and noun-noun compounds are single tokenized on the target side and named entities are aligned on both sides. It may be observed that similar treatments on the source and the target sides separately have achieved best performance of the system in the respective cases.

The system performance improves when the alignment list of complex predicates is incorporated in the baseline best system. When this system is augmented with single tokenization of phrasal prepositions and verb-object combinations on the source side and single tokenization of reduplicated phrases and noun-noun compounds on the target side, it achieves the BLEU score of 15.12. This is the best result obtained so far with respect to the baseline system (6.38 BLEU points absolute, 73% relative improvement). It may be observed from Table 2 that baseline Moses without any preprocessing of the dataset produces a BLEU score of 8.74.

Intrinsic evaluation of the word alignment could not be performed as gold-standard word alignment was not available. 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). Bengali is a morphologically rich language and has relatively free phrase order. Proper evaluation of the English-Bengali MT evaluation ideally requires multiple set of reference translations. Moreover, the training set was smaller in size.

It is observed in the evaluation results that if noun-noun compounds are treated as single tokens on both sides, the score is reduced. The various reasons that have been identified for this observation are: (i) noun-noun compounds on the source side may not be treated as noun-noun compounds on the target side, (ii) number of tokens on both sides do not match, (iii) wrong identification of some noun-noun compounds on the source side and (iv) most of the noun-noun compounds on the target side are named entities which have been already identified in the baseline best system.

6 Conclusions and Future work

We have presented a system to show how the simple yet effective preprocessing of various types of MWEs can boost the performance of PB-SMT system on an English—Bengali translation task. Our best system yields 6.38 BLEU points improvement over the baseline, a 73% 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 almost always looks better in terms of either lexical choice or word ordering. It is observed that only 28.5% of the test set NEs appear in the training set, yet prior automatic alignment of the NEs and complex predicates improves the translation quality. This suggests that not only the NE alignment quality in the phrase table but also the word alignment and phrase alignment quality improves significantly. At the same time, single-tokenization of MWEs makes the dataset sparser, but improves the quality of MT output to some extent. Data-driven approaches to MT, specifically for scarce-resource language pairs for which very little parallel texts are available, should benefit from these preprocessing methods. Data sparseness is perhaps the reason why single-tokenization of NEs and compound verbs, both individually and in collaboration, did not add significantly to the scores. However, a significantly large parallel corpus can take care of the data sparseness problem introduced by the single-tokenization of MWEs.

The present work offers several possibilities for further work. We will investigate the role of noun-noun compounds and their alignment for improving the overall performance of the MT system. Moreover, identification of the various types of MWEs on both sides needs improvement. Further work will be carried out by considering alignment lists of the various types of MWEs for which single tokenization has been done.

Acknowledgement

The work has been carried out with support from the consortium-mode project “Development of English to Indian Languages Machine Translation (EILMT) System funded by Department of Information Technology, Government of India.

References

Agarwal, Aswini, Biswajit Ray, Monojit Choudhury, Sudeshna Sarkar and Anupam Basu. 2004. Automatic Extraction of Multiword Expressions in Bengali: An Approach for Miserly Resource Scenario. *In Proc. of International Conference on Natural Language Processing (ICON)*, pp. 165-174.

Baldwin, Timothy and Su Nam Kim (2010) *Multiword Expressions,* in Nitin Indurkhya and Fred J. Damerau (eds.) *Handbook of Natural Language Processing*, Second Edition, CRC Press, Boca Raton, USA, pp. 267—292.

Banerjee, Satanjeev, and Alon Lavie. 2005. An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *In proceedings of the ACL-2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization,* pp. 65-72. Ann Arbor, Michigan., pp. 65-72.

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. *The mathematics of statistical machine translation: parameter estimation. Computational Linguistics*, 19(2):263-311.

Carpuat, Marine, and Mona Diab. 2010. Task-based Evaluation of Multiword Expressions: a Pilot Study in Statistical Machine Translation. *In Proc. of Human Language Technology conference and the North American Chapter of the Association for Computational Linguistics conference (HLT-NAACL 2010)*, Los Angeles, CA.

Chakrabarti, Debasri, Hemang Mandalia, Ritwik Priya, Vaijayanthi Sarma, and Pushpak Bhattacharyya. 2008. Hindi compound verbs and their automatic extraction. *In Proc. of the 22nd International Conference on Computational Linguistics (Coling 2008)*, Posters and demonstrations, Manchester, UK, pp. 27-30.

Chakraborty, Tanmoy and Sivaji Bandyopadhyay. 2010. Identification of Reduplication in Bengali Corpus and their Semantic Analysis: A Rule Based Approach. *In proc. of the 23rd International Conference on Computational Linguistics (COLING 2010), Workshop on Multiword Expressions: from Theory to Applications (MWE 2010).* Beijing, China.

Das Dipankar, Santanu Pal, Tapabrata Mondal, Tanmoy Chakraborty, Sivaji Bandyopadhyay .2010. Automatic Extraction of Complex Predicates in Bengali *In proc. of the workshop on Multiword expression: from theory to application (MWE-2010)*, The 23rd International conference of computational linguistics (Coling 2010),Beijing, Chaina, pp. 37-46.

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): 1–38.

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

Eck, Matthias, Stephan Vogel, and Alex Waibel. 2004. Improving statistical machine translation in the medical domain using the Unified Medical Language System. *In Proc. of the 20th International Conference on Computational Linguistics (COLING 2004)*, Geneva, Switzerland, pp. 792-798.

Ekbal, Asif, and Sivaji Bandyopadhyay. 2008. Maximum Entropy Approach for Named Entity Recognition in Indian Languages. *International Journal for Computer Processing of Languages (IJCPOL)*, Vol. 21 (3), 205-237.

Ekbal, Asif, and Sivaji Bandyopadhyay. 2009. Voted NER system using appropriate unlabeled data. *In proc. of the ACL-IJCNLP-2009 Named Entities Workshop (NEWS 2009)*, Suntec, Singapore, pp.202-210.

Feng, Donghui, Yajuan Lv, and Ming Zhou. 2004. A new approach for English-Chinese named entity alignment. *In Proc. of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004),* Barcelona, Spain, pp. 372-379.

Huang, Fei, Stephan Vogel, and Alex Waibel. 2003. Automatic extraction of named entity translingual equivalence based on multi-feature cost minimization. *In Proc. of the ACL-2003 Workshop on Multilingual and Mixed-language Named Entity Recognition, 2003*, Sapporo, Japan, pp. 9-16.

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, pp. 181–184. Detroit, MI.

Koehn, Philipp, 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.

Koehn, Philipp, 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.

Koehn, Philipp. 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 388-395.

Moore, Robert C. 2003. Learning translations of named-entity phrases from parallel corpora. *In Proc. of 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003)*, Budapest, Hungary; pp. 259-266.

Mukherjee, Amitabha, Soni Ankit and Raina Achla M. 2006. Detecting Complex Predicates in Hindi using POS Projection across Parallel Corpora. *Multiword Expressions: Identifying and Exploiting Underlying Properties Association for Computational Linguistics,* pp. 28–35, Sydney.

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.

Pal Santanu, Sudip Kumar Naskar, Pavel Pecina, Sivaji Bandyopadhyay and Andy Way.2010. Handling Named Entities and Compound Verbs in Phrase-Based Statistical Machine Translation, *In proc. of the workshop on Multiword expression: from theory to application (MWE-2010)*, The 23rd International conference of computational linguistics (Coling 2010),Beijing, Chaina, pp. 46-54.

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.

Paul, Soma. 2010. Representing Compound Verbs in Indo WordNet. *Golbal Wordnet Conference-2010*, pp. 84-91.

Rayson, Paul, Dawn Archer, Scott Piao, and Tony McEnery. 2004. The UCREL Semantic Analysis System. *In proc. Of LREC-04 Workshop: Beyond Named Entity Recognition Semantic Labeling for NLP Tasks*, pages 7-12, Lisbon, Porugal.

Ren, Zhixiang, Yajuan Lü, Jie Cao, Qun Liu, and Yun Huang. 2009. Improving statistical machine translation using domain bilingual multiword expressions. *In Proc. of the 2009 Workshop on Multiword Expressions, ACL-IJCNLP 2009*, Suntec, Singapore, pp. 47-54.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. *In Proc. of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002),* Mexico City, Mexico, pp. 1–15.

Sinha, R. Mahesh, K. 2009. Mining Complex Predicates In Hindi Using A Parallel Hindi-English Corpus. *Multi Word Expression Workshop, Association of Computational Linguistics-International Joint Conference on Natural Language Processing-2009*, pp. 40-46, Singapore.

Tanaka, Takaaki and Timothy Baldwin. 2003. Noun-Noun Compound Machine Translation: A Feasibility Study on Shallow Processing. In *Proc. of the Association for Computational Linguistics- 2003, Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan, pp. 17–24.

Venkatapathy, Sriram, and Aravind K. Joshi. 2006. Using information about multi-word expressions for the word-alignment task. *In Proc. of Coling-ACL 2006: Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties, Sydney*, pp. 20-27.

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.

Wu, Hua Haifeng Wang, and Chengqing Zong. 2008. Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. *In Proc. of the 22nd International Conference on Computational Linguistics (COLING 2008),* Manchester, UK, pp. 993-1000.