Event and Event Actor Alignment in Phrase Based Statistical Machine Translation

Anup Kumar Kolya[1], Santanu Pal[1]
[1]Dept. of Computer Science & Engineering
Jadavpur University
Kolkata-700 032, India
{anup.kolya, santanu.pal.ju}gmail.com

Asif Ekbal[2], Sivaji Bandyopadhyay[1]
[2]Dept. of Computer Science & Engineering
IIT Patna,
Patna-800 013, India
asif@iitp.ac.in, sivaji_ju_cse@yahoo.com

Abstract

This paper proposes the impacts of event and event actor alignment in English and Bengali phrase based Statistical Machine Translation (PB-SMT) System. Initially, events and event actors are identified from English and Bengali parallel corpus. For events and event actor identification in English we proposed a hybrid technique and it was carried out within the TimeML framework. Events in Bengali are identified based on the concept of complex predicate structures. There can be one-to-one and one-to-many mappings between English and Bengali events and event actors. We preprocess the parallel corpus by single tokenizing the multiword events and event-actors which reflects some significant gain on the PB-SMT system. We represent a hybrid alignment approach of events and event-actors in both English-Bengali training corpus by defining a rule based aligner and a statistical hybrid aligner. The rule base aligner assumes a heuristic that the sequence of events and event actors on the source (English) side are also maintained in the target (Bengali) side. The performance of PB-SMT system could vary depending on the number of events and event-actors that are identified in the parallel training data. The proposed system achieves significant improvements (5.79 BLEU points absolute, 53.02% relative improvement) over the baseline system on an English-Bengali translation task.

1 Introduction

Event and event actor alignment play a very crucial role to improve the translation quality in a machine translation system. A translated sentence is not a satisfactory and proper translation until we properly combine event and event actor in sentence level task. Recently, event related works are becoming popular in the machine translation field. Sentence-aligned parallel bilingual corpora are very useful for applying machine learning approaches to machine translation. But, most of these works have been focused on European language pairs and some of the Asian Languages such as English-Japanese and English-Chinese. In this work, we have added event and event-actor alignments as additional parallel examples with the English-Bengali parallel corpus. The entire task is divided into three steps, first, we identify event and event actors on the both side of the parallel corpus, second, we align events and event actors using a rule based and a statistical alignment method and finally, the identified multiword events and event actors are single tokenized on the both side and then the prior alignment of event and event actors are applied on the English-Bengali PB-SMT system for further improvement.

The identification of events on English side, we have followed the guidelines of TimeML view (Pustejovsky et al., 2003a). TimeML defines events as situations that *happen or occur*, or elements describing *states or circumstances* in which something obtains or holds the truth. These events are generally expressed by tensed or un-tensed verbs, nominalizations, adjectives, predicative clauses or prepositional phrases. In the sentences, almost all events are involved with the event actor, either active or passive. Event actor identification in English is facilitated by the available free resources and tools such as Stanford Parser, VerbNet (Kipper-Schuler et al, 2005) .In detail research works related to English event and event actor identification can be found in (Kolya et al., 2010).

We have defined Complex Predicates as events (Das et al., 2010) in Bengali. Complex Predicates (*CPs*) in Bengali consists of both compound verbs and conjunct verbs. Complex Predicates contain [*verb*] + *verb* (*compound verbs*) or [*noun/ adjective/adverb*] +verb (*conjunct verbs*) combinations in *South Asian language*s (Hook, 1974).

In the next step, we identify event actors of event from Bengali language. We have

International Joint Conference on Natural Language Processing, pages 36–44,
Nagoya, Japan, 14-18 October 2013.

considered the same guidelines for event actor identification in Bengali as those proposed for event actor identification in English. For Bengali event actor identification, we have used two available lexical engines, namely Name Entity Recognizer (NER) (Ekbal and Bandyopadhyay, 2009) and shallow parser[1]. The accuracy of the Bengali NE recognizer (NER) is poorer compared to English NER because (i) there is no concept of capitalization in Bengali (ii) some Bengali common nouns are also often used as named entities. Similarly, the Bengali shallow parser faces such kinds of difficulties. Overall, Bengali is morphologically rich language and has very limited such kind of resources.

The major challenge is to develop an event alignment system between a resource-rich language like English and a resource-poor language like Bengali. The proposed system is relying on the design of rules and the availability of large amounts of annotated data. But, building of large amount data is a time consuming, labour intensive and expensive task.

The main motivation of this work is the scarcity of sufficient works related to event alignment. To the best of our knowledge this is the first time that the event alignment approach is applied for the English-Bengali language pair. Given a set of parallel sentences, we identify events and event actors in both the sides. The events and event actors in both sides of the parallel corpus are assigned appropriate tags (event: e and event actor: ea). Thereafter we align the English events and event actors with Bengali events and event actors. The alignment has been carried out by single tokenizing the multi word events and event-actors on both sides of the parallel corpus. Thereafter the alignment of events and event actors in the parallel English-Bengali sentences is carried out based on two approaches: (i) rule based approach and (ii) hybrid statistical approach. The rule based approach fails to align the causal sentences that include the cause-effect constructs. The positions of the cause and the effect clauses may change their position in the target sentence. The positions of the cause and the effect clauses may change their position in the target sentence. Such types of parallel sentences are event aligned using the hybrid statistical approach. We attempt to achieve good accuracies for event identification and event actor identification for both the languages which is reflected as the improvement of the English-Bengali PB-SMT system performance. The hybrid approach also validates the correctness of the alignment of the rule based system.

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

The works related to alignment are mostly developed for machine translation task. Some works in sentence alignment can be found in (Brown, 1991) and (Gale and Church, 1993). (Chen, 1993) developed a method which was slower but more accurate than the sentence-length based Brown and Gale algorithm. (Wu, 1994) used an approach which was adopted from Gale and Church's method for Chinese. They used a small corpus-specific bilingual lexicon to improve alignment accuracy in texts containing multiple sentences of similar length. (Melamed 1996, 1997) also proposed a method based on word correspondences. (Plamondon, 1998) developed a two-pass approach, in which a method similar to the one proposed by Melamed identifies points of correspondence in the text that constrain a second-pass search based on the statistical translation model. (Moore, 2002) developed a hybrid sentence-alignment method using sentence length-based and word-correspondence-based models. This model is fast, very accurate, and requires that the corpus be separated into words and sentences. In the hybrid model, they used the sentence pairs that are assigned the highest probability of alignment to train a modified version of IBM Translation Model 1 (Brown, 1993). (Fung, 1994) presented K-vec, an alternative alignment strategy, that starts by estimating the lexicon. 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

[1] http://ltrc.iiit.ac.in/showfile.php?filename=downloads/shallow_parser.php

automatically extracting NE translingual equivalences between Chinese and English based on multi-feature cost minimization has been proposed in Huang et al. (2003).

3 System Description

In our system, initially we have identified Event and Event Actor from English-Bengali parallel corpus. Then, we have established Rule base event and event-actor Alignment Model, and Statistical Hybrid based Alignment model for the experiment setup.

3.1 English Event Identification

Our approach for event identification is based on a hybrid approach. The system is combined with Support Vector Machine (SVM[2, 4]), semantic role labeling (SRL) (Gildea et al, 2002; WorldNet[7] and several heuristics.

Hybrid event identification system

Some lexical rules have been used to identify the de-verbal event words more accurately, in addition with SVM, SRL, WordNet based approaches. Rules are extracted on the basis of detailed analysis of suffixes and the morphological markers of de-verbal derivations like *'expedition*' and '*accommodation'* in the source side of the corpus. Initially, Stanford Named Entity (NE) tagger[3] is passed on the English side of the training corpus. The output of the system is tagged with *Person*, *Location*, *Organization* and *Other* classes. The following cue sets or rules are applied for event extraction:
Cue-1: The morphologically de-verbal nouns are usually identified by the suffixes like '*-tion*', '*-ion*', '*-ing*' and '*-ed*' etc. The non-NE nouns that end with these suffixes are considered as the event words.
Cue 2: After searching verb-noun combination from the test set, non-NE noun words are considered as the events.
Cue 3: The non-NE nouns occurring after (i) the complements of aspectual PPs headed by prepositions, (ii) any time-related verbs and (iii) certain expressions are considered as events.

The performance of the event extraction system has been reported with the precision, recall and F-measure values of 93.00%, 96.00% and 94.47%, respectively on the TempEval-2 corpus.

3.2 Event-Actor identification

It has been observed from the detailed text analysis that almost all events are associated with some actors ("*anything having existence (living or nonliving)"*), either active or passive. More generally, event actions are associated with persons or organizations and sometimes with locations. In this section, it has been shown how event actors are identified for the events.

Subject Based Baseline Model

The input English sentences with event constructs are passed through the Stanford Parser to extract the dependency relationships from the parsed data. The output is checked to identify the predicates, "*nsubj*" and "*xsubj*" so that the *subject* related information in the "*nsubj*" and "*xsubj*" predicates are considered as the probable candidates of event actors. Other dependency relations are filtered out.

Syntax Based Model

The syntax of a sentence in terms of its argument structure or sub-categorization information of the associated verb plays an important role to identify the event actors of the events in a sentence.

(a) Syntax Acquisition from Verbnet

Using VerbNet (Kipper-Schuler et al, 2005), a separate rule based argument structure acquisition system is developed in the present task for identifying the event actors. The acquired argument structures are compared against the extracted VerbNet frame syntaxes. If the acquired argument structure matches with any of the extracted frame syntaxes, the event actor corresponding to each event verb is tagged with the actor information in the appropriate slot in the sentence.

(b) Argument Structure Acquisition Framework

To acquire the argument structure, Stanford Parser parsed event sentences are passed through a rule based *phrasal-head* extraction system to identify the *head part* of the phrase (well-structured and bracketed) level argument structure of the sentences corresponding to the event verbs.

SRL for Event Actor Identification

Semantic Role Label (SRL) plays an important role to extract target argument relationship from

[3]http://chasenorg/~taku/software/yamcha
[4]http://nlp.stanford.edu/software/CRF-NER.shtml

the semantic role labeled sentences. Here, the argument is considered as an event actor and the target is identified as the corresponding event. Let us consider the following example:

*[**ARG1** A military coup] [**TARGET** followed], during which [**ARG1** Allende] [**TARGET** committed] suicide rather than surrender to his attackers.*

In the first trace, ***[A military coup]*** is identified as the event actor <eActor> of the corresponding event word *[**followed**]*. In the second trace, ***[Allende]*** is the event actor <eActor> of the corresponding event ***[committed]***. So using the SRL technique, the event and the corresponding event actor are found. The original F-scores of the event actor identification systems for the subject based and syntax based models are 65.98% and 70%, respectively. Adding the SRL technique for event actor identification, the F-score of the system further improves to **73%**.

3.3 Bengali Event Extraction

The sentences are passed through an open source Bengali shallow parser[1]. The shallow parser gives different morphological information (root, lexical category of the root, gender, number, person, case, etc.) that helps in identifying the lexical patterns of Complex Predicates (*CPs*).

Bengali sentences were POS-tagged using the available shallow parser. We have extracted{verb(v)+verb(v), (noun(n)+verb(v)) and (adjective(adj)+verb(v))} lexical complex predicates pattern. The complex predicate (v+v) pattern is considered as the compound verb and (n+v) and (adj+v) patterns are considered as conjunct verbs (*ConjVs*). These compound and conjunct verb patterns are used as the possible candidates for event expressions.

Identification of Complex Predicates (CPs)

In the Bengali side, generally complex predicates follow some patterns such as conjunct verbs (e.g., মেরে ফেলা [*mere phela*] 'to kill'): adjective/adverb/noun +verb pattern or compound verbs (e.g., ভরসা করা [*bharsha kara*] 'to depend'): verb + verb pattern. To identify such complex Predicates (CPs) in Bengali, Morphological knowledge is required. Compound verbs consist of two verbs – a full verb followed by a light verb. The full verb is represented either as conjunctive participial form -এ [–*e*] or the infinitive form -তে [–*te*] at the surface level. The light verb bears the inflection based on tense, aspect and person information of the subject. On the other hand, each Bengali conjunct verb consists of adjective, adverb or noun followed by a light verb. These light verbs are semantically lightened, polysemous and limited into some definite candidate seeds (Paul, 2010).

The other types of predicates presents in Bengali language follow the same lexical pattern like the compound verb but the Full Verb and Light Verb behave as independent syntactic entities (e.g, নিয়ে গেল niye gelo 'take-go'). Such complex predicates are termed as Serial Verb (SV).

Das et al. (2010) analyzed and identified the categories 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.

3.4 Bengali Event Actor Identification

Here, events are associated with either active or passive event actors in Bengali like in English language. Similarly, event actions are associated with persons or organizations and sometimes with locations. Initially, sentences that do not have any event words are discarded.

Bengali Name Entity Recognizer (NER) and Bengali shallow parser are employed to detect the event actors from the sentences. The baseline system for identifying event actor is developed based on the person, organization and location information which are recognized by Bengali NER. Then, Bengali shallow parser has been used to improve the performance of event actor identification. In the following two sections, it has been shown in details how event actors are identified for the events in Bengali language by applying the above two techniques.

Name Entity based Approach

Here, Bengali named entities are identified from parallel corpus. After identification of Bengali NEs and Bengali events from the sentences, following heuristics rules are introduced for event actor identification:

(i) If sentence is having only one NE and one or more than one events then this single NE is selected as the event actor for all events.

(ii) If sentence is having multiple NEs and only one event, then all the NEs are selected as the event actors for the single event.

(iii) If there exists multiple NEs and multiple events in a sentence, then <event, actor> pairs are formed by considering an event and its closest possible NE as the event actor in the sentence.

Example: <ea> <**আন্টার্টিকা**>**</ea> <e>পরিবর্তিত হচ্ছে</e>**, সেটা প্রাকৃতিক না মানুষের জন্য এই পরিবর্তন.
<ea>শাসকের **</ea>**>দ্বারা<e> **অত্যাচারিত হয়ে</e>**.

Shallow Parsing approach

Bengali pronouns (PRP) are not identified by the Bengali NER. The shallow parser is used to identify the pronouns in a sentence that can play the role of event-actor of event. Initially, the input Bengali sentences are passed through the shallow parser to extract phrase and POS information from the parsed data. Here, noun phrases (NP) and verb phrases (VP) are only considered from the parsed output. From noun phrases, the word with the pronoun (PRP tag) is extracted as the event actor of the corresponding event expressed in the verb phrase (VP).

<ea>**যারা**/PRP/NP **</ea>** *সাক্ষাত/*NN/NP *করেছিল/*VM/VP আর্টলান্টিক/JJ/NP <ea>**তাদের**/PRP/NP</ea> *মনে* NN/NP *করিয়ে*/VM/VGF *দেয়* /VAUX/VGF প্রকৃতির/NN/NP ভয়ঙ্কর/JJ/NP সন্ত্রস্ত/JJ/JJP

3.5 Rule based event and event-actor Alignment Model

The rule based alignment model aligns the identified events and event-actors between the English and Bengali parallel sentences. Here it is observed that that event-actors associated with events appear as contiguous sequence of words in a sentence. For example, “*travelers”* is an event actor of the event word *“discover”* in the English side which is aligned with “ভ্রমণকারী”, the event actor of the event word <**আবিষ্কার করবে**> in the Bengali side. “Discover” is an event group with the syntactic structure event actor *“travelers”* which can be determined deterministically given the phrase (NP, VP) and POS tags information.

Ex-(a) ...adventurous/JJ travelers/NNS will/MD discover/VB an/DT ethereal ……
Ex-(b) ...দুঃসাহসিক ভ্রমণকারী <**আবিষ্কার করবে**> একটা.....

During event and event actor alignment the following issues are observed between the English and the Bengali language:
(i) It aligns both one-to-one and one-to-many alignments between word forms.
(ii) In the English and Bengali side event actors are identified by noun (NN), proper noun (NNP) and pronoun (PRN) based word from the noun phrase. Then the alignment has been done on both sides.
(iii) In event alignment, English side event words are generally verb(VB) and noun(NN) while the internal structure of Bengali event words are combination of compound verbs (VM-Vaux) and conjunct words (NN-VAUX,ADJ-VUX).
(iv) In event alignment, English event words are generally aligned to a group of Bengali event words. Light verbs are added with the main verb which increases the number of words in Bengali with respect to English event word in the sentence. Similarly for English event words, the auxiliary verb is considered as a part of it. The following alignment from Example (a) above bears testimony to the above.

will discover **→** ***আবিষ্কার করবে.*** *[abiskar korbe]*

Example 2: *Adventurous <ea>* ***traveler*** *</ea> will<event>* ***discover****</event> an ethereal landscape that <event>* ***lingers*** *</event> in the memory.*

দুঃসাহসিক *<ea>* **ভ্রমণকারী** *</ea> <event>* **আবিষ্কার করবে** *</event>* একটা স্বর্গীয়স্থান যেটা *<event>* **মনে রাখার** *</event>*মতো.

In the above parallel sentence, the event actor “*traveler*” on the English side is aligned with “*ভ্রমণকারী*” on the Bengali side. The corresponding events associated with the event actor are “*discover*” and “lingers” on the English side which are aligned with “ *আবিষ্কার করবে*” and “ *মনে রাখার* ” respectively in the Bengali side. In order to get the correct alignment, identification of event actors and events orders should be correct. Thus the following parallel phrase translation entries are generated.

Traveler ↔ **ভ্রমণকারী [vramonkari]**
will discover ↔**আবিষ্কার করবে [abiskar korbe]**
Lingers ↔**মনে রাখার [mone rakhar]**

(v). It has been observed that the order of event actor with event in English and Bengali language are same in most of the cases. Correct identification of event words in Bengali side corresponding to English side plays an important role in the event word alignment. In the example 2, it is easy to align, but in some cases the word alignment complexity increases when the order of the events and the event actors does not follow the same sequence in the English and the Bengali parallel sentences.

The complexity is further increased due to the non-availability of large bilingual corpus and the presence of inflectional variations in Bengali. So sometimes it is difficult to correctly align event words to the target words. Once these alignments are obtained, then we validate the alignment with statistical hybrid based alignment model.

3.6 Hybrid based Alignment model

Initially an English-Bengali phrase based statistical translation model has been developed which has been trained with the same EILMT tourism domain corpus of 22,492 sentences. The above rule based event actor alignments are validated by translating both the event and the event actor. From the above knowledge we get a link between the event and the event actor on both sides. Even the alignment details are also available. . From this point of view, we can conclude that if we know any of the translation of either the event or the event actor then we can align with the target event and event actor relation. Using this heuristics, we have translated event or event actor and matched with the target Bengali event or event actor which has been provided by the rule based system as described in section 4. A string level edit distance matric has been used to validate the bilingual even-actor relations. After alignment of event and actor words from English side, we collect token position number of the event words with event tag from the sentence. We follow the Timex3 guideline for event word identification, so English side event words are mainly single word based token. Position of the single token number is added with event tag **<*e*>**. For the identification of event actors in the Bengali side, we follow the guidelines of English event actor **<*ea*>** identification that is already defined in Rule no (ii) in section 4. On English side after identification of event word in a sentence, we have added auxiliary dependent verb with it as defined in rule no (iv).

After identification we have pre-processed the single tokenized corpus by replacing space with underscore (‘_’). We have used underscore (‘_’) instead of hyphen (‘-’) because there already exists some hyphenation words in the corpus. The use of Underscore (‘_’) character also facilitates the de-tokenizing the single-tokenized events or event-actors at decoding time.

Amidst[0] such[1] solitude[2], adventurous[3] **<*ea*> *travelers[4]* </*ea*>** *will[5]***<*e*> *discover[6]*</*e*>** *an[7] ethereal[8] landscape[9] that[10]* **<*e*> *lingers[11]* </*e*>** *in [12]the[13] memory[14].*

After considering depending auxiliary verb
Amidst[0] such[1] solitude[2], adventurous[3] **<*ea*> *travelers[4]* </*ea*> <*e*> *will_discover[5]*</*e*>** *an[6] ethereal[7] landscape[8] that[9]* **<*e*> *lingers[10]* </*e*>** *in [11]the[12] memory[13].*

দুঃসাহসিক*[0] <ea>* ভ্রমণকারী***[1]****</ea> <e>* **আবিষ্কার_করবে*[2]*** *</event>* একটা*[3]* স্বর্গীয়স্থান*[4]* যেটা*[5] <event>* **মনে_রাখার *[6]***</event>*মতো.

We collect the token position number of event word(s) and actor(s) from both sides of the parallel sentence. Finally we get a sentence level source-target event-event actor-actor alignment.

For example, 4-1 5-2 11-6

We have also generated source-target event and event-actor alignment level parallel example which has been added as additional parallel example with the training data. Now we retrain the PB-SMT system using moses toolkit (Koehn et at., 2003). The sentence level positional alignment information helps us for updating and correcting the alignment table which has been generated during the training phase using grow-diag-final-and algorithm. The rest of the process has been followed as described in the state-of-art system.

This approach also helps us to align the event and event-actor relation which cannot be aligned by the rule based system. In this approach we have translated the identified source events or event-actors. The translated events or event-actors are matched with the corresponding target side events and event-actors by using string level edit-distance method.

4 Tools and Resources

A sentence-aligned English-Bengali parallel corpus containing 23,492 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[4]”. The Stanford Parser[5],

[4] The EILMT project is funded by the Department of Electronics and Information Technology (DEITY), Ministry of

Stanford NER, CRF chunker[6] and the Wordnet 3.0[7] have been used for identifying the events and the event-actors 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[8]".

The effectiveness of the present work is demonstrated by using the standard log-linear PB-SMT model as our baseline system. The 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) have been used in the present study.

5 Experiments and Evaluations

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 22,492 sentences. In addition to the target side of the parallel corpus, a monolingual Bengali corpus containing 488,026 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 7 produce the optimum baseline result. The baseline model (Experiment 1) has scored 10.92 BLEU matric points that is described in Table 3. We carried out the rest of the experiments using these settings. Initially we identified event actor relation on both sides of the parallel corpus by developing an automatic Event actor Identifier. The system achieves Recall, Precision and F-Score values of 82.06%, 72.32% and 75.73% respectively for Bengali event identification in training corpus.

In the Bengali event actor evaluation framework, we have randomly selected 500 sentences from the Bengali corpus for testing. Each sentence is having around maximum100 words. We have manually annotated these 500 sentences with event actor tag as the reference data. The evaluation results for Bengali event-actor identification in the training corpus are shown in Table 1.

Type	**Baseline Model**	**Combination of NER and Shallow Parser Model**
Precision	51.31	58.12
Recall	56.74	55.90
F-measure	53.89	56.99

Table 1: Evaluation results of Bengali event actor identification

Training set	English		Bengali	
	T	U	T	U
Event	8142	3889	20174	7154
Actor	21931	12273	17107	11106

Table 2: Event and Event-actor Statistics (T - Total occurrence, U – Unique)

Table 2 shows the statistics of events and event actors in the English and Bengali corpus. In the training corpus, 44.5% and 47.8% of the event actors are single-word event actors in English and Bangla respectively, which suggests that prior alignment of the single-word event actors, in addition to multi-word event actors alignment, should also be beneficial to word and phrase alignment.

Our experiments have been carried out in three directions (i) Initially we single tokenized the identified events and event-actors on both sides of the parallel corpus (ii) we added the single tokenized event and event-actor alignment as an additional parallel data with the training corpus and (iii) we updated the word alignment table using hybrid word alignment technique. The table 3 shows that the successive evaluation of different experimental settings of PB-SMT system. Experiment 1 reports the baseline model score of the PB-SMT system. In experiment 2, we preprocessed the parallel corpus by single tokenizing the events and event actors, this

Communications and Information Technology (MCIT), Government of India.

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

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

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

[8] The IL-ILMT project is funded by the Department of Electronics and Information Technology (DEITY), Ministry of Communications and Information Technology (MCIT), Government of India.

makes significant gain over baseline system. Rest of the experiments (3, 4, 5 and 6) has been carried out with single tokenization of event and event actors along with their alignments. Experiment 3 and 4 reports that the alignment of events and event actors are added with the parallel corpus also improve the MT system performance. In experiment 5, both event and event actor alignments are combined together as additional parallel data with the training corpus, produced 5.51 (50.45%) BLEU point relative improvement over the baseline system. While in experiment 6, we updated the alignment table using event and event-actor alignment the performance has increased significantly with 5.79 (53.02%) BLEU point relative improvement over baseline system.

Experiments		No.	BLEU	NIST
Baseline		1	10.92	4.13
Single tokenized Event and Event-Actor		2	12.68	4.33
Experiment 2	Event actor alignment as additional parallel data	3	15.23	4.47
	Event alignment as additional parallel data	4	13.48	4.37
	Event and event actor alignment as additional parallel data	5	16.43	4.51
	event actor alignment (by updating word alignment table) †	6	**16.71**	**4.54**

Table 3: Evaluation results (The '†' marked systems produce best score)

6 Conclusions and Future work

The present work shows how three approaches (i) single tokenization of event and event-actors on both sides of the parallel corpus (ii) alignment of event and event-actor added as an additional training data with the parallel corpus and (iii) updating the word alignment table directly by event-actor and event alignment boost up the performances of the overall system. The method also reduces data sparsity problem. The single tokenization helps us to bound multi word events and event-actors into a single unit. On manual inspection we see that the translation output looks better than the baseline system output in terms of better lexical choice and word ordering. On experiment 3 and 4 our systems achieve 5.51 BLEU points absolute, 50.45% and 5.79 BLEU points absolute, 53.02% relative improvement over the baseline system on an English-Bengali translation task. The event and event actor alignment performance is also reflected indirectly by increasing the MT performance. The fact that only 28.5% of the testset event-actors appear in the training set, yet prior automatic alignment of the event and event actors brings about so much improvement in terms of MT quality, suggests that it not only improves the event and event actor alignment quality in the phrase table, but word alignment and phrase alignment quality must have also been improved significantly.

Our future work will be focused on post editing the MT output using event and event-actor relation. As event and event-actor plays an important role in terms of discourse, we can reorder the output target sentences according to the occurrences of event on the source side. We will also focus to upgrade our system for paragraph translation. In future we can add temporal expression and location of event with event-actor as attributes. These attributes of event can further improve the performance machine translation result.

Acknowledgement

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

References

Brown, P.F., Della Pietra, S. A., Della Pietra, V. J., Mercer, R.L.(1993). *The Mathematics of Statistical Machine Translation:* Parameter Estimation. Computational Linguistics 19(2) 263–311.

Brown, P.F., Lai, J.C. and Mercer, R.L. (1991). *Aligning Sentences in Parallel Corpora*. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics,Berkeley, California 169–176.

Chen, S.F.: 1993. *Aligning Sentences in Bilingual Corpora Using Lexical Information*. In Proceedings of the 31st Annual Meeting of the ACL, Columbus, Ohio (1993) 9–16.

Das,D., Pal,S. Mondal,T. Chakroborty,T. and Bandyopadhyay,S.:*Automatic Extraction of Complex Predicates in Bengali* . MWE 2010 Workshop, Coling 2010, Beijing, China.

Ekbal, A. and Bandyopadhyay,S.(2009).*"Voted NER system using appropriate unlabeled data*". In proceedings 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.

Fung,P. and CHURCH, K.: “K-vec.(1994). *A New Approach for Aligning Parallel Texts*. In COLING-94: 15th International Conference on Computational Linguistics, Kyoto: Aug., 1096--1102.

Gale,W.A., Church, K.W.: *A program for Aligning Sentences in Bilingual Corpora*. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, California (1991) 177–184.

Gildea, D. and Jurafsky, D. (2002). *Automatic Labeling of Semantic Roles*. Computational Linguistics, 28(3):245–288 .

Hook, P. (1974). *The Compound Verbs in Hindi*. The Michigan Series in South and South-east Asian Language and Linguistics. The University of Michigan.

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.

Kipper-Schuler and K.: VerbNet.(2005). *A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis,Computer and Information Science Dept., University of Pennsylvania, Philadelphia,PA .

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.

Kolya, A. Das, D. Ekbal A. and Bandyopadhyay, S.(2011). *A Hybrid Approach for Event Extraction and Event Actor*. In RANLP,12-14 September, Hissar, Bulgaria PP.592-597.

Melamed, I.D.(1996). *A Geometric Approach to Mapping Bitext Correspondence*. IRCS Technical Report 96-22, University of Pennsylvania.

Melamed, I.D.(1997).*A Portable Algorithm for Mapping Bitext Correspondence*. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, Madrid, Spain 305–312

Moore, Robert C. (2002), *Fast and Accurate Sentence Alignment of Bilingual Corpora*.AMTA, 135-144.

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.

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.

Paul, S. (2004). *An HPSG Account of Bangla Compound Verbs with LKB Implementation.* Ph.D dissertation, University of Hyderabad, Hyderabad.

Pustejovsky,J., Castano,J., Ingria, R.Sauri,R., Gaizauskas,R., Setzer,A., Katz, and Radev.(2003). *TimeML: Robust Specification of Event and Temporal Expressions in text.* In Proceedings of the Fifth International Workshop on Computational Semantics (IWCS-5), Tilburg.

Stolcke, A. SRILM—An Extensible Language Modeling Toolkit. Proc. Intl. Conf. on Spoken Language Processing, vol. 2, pp. 901–904, Denver (2002).