Hybrid Word Alignment

Santanu Pal and Sudip Kumar Naskar

Abstract This paper proposes a hybrid word alignment model for Phrase-Based Statistical Machine Translation (PB-SMT). The proposed hybrid word alignment model provides most informative alignment links, which are offered by both unsupervised and semi-supervised word alignment models. Two unsupervised word alignment models, namely GIZA++ and Berkeley aligner, and a rule based word alignment technique are combined together. The unsupervised alignment models are trained on the surface form as well as the root form of the training data and provide alignment tables for the corresponding training data. The rule-based aligner is aimed towards aligning named entities (NEs) and syntactically motivated chunks. NEs are aligned through transliteration using a joint source-channel model. Chunks are aligned employing a bootstrapping approach by translating the source chunks into the target language using a baseline PB-SMT model and subsequently validating the chunk hypotheses using a fuzzy matching technique against the target corpus. Experiments are carried out after single-tokenizing the multiword NEs. The effectiveness of the proposed hybrid alignment model was extrinsically evaluated on the MT quality by using well-known automatic MT evaluation metrics, such as BLUE and NIST. Our best system provided significant improvements over the baseline as measured by BLEU.

1 Introduction

Word alignment is the backbone of PB-SMT systems or any data driven approaches to Machine Translation (MT) and it has received a lot of attention in the area of statistical machine translation (SMT) (Brown et al. 1993; Och et al. 2003; Koehn et al. 2003) as the success of SMT or any other data driven approaches to MT is

S. Pal (✉)
Universität Des Saarlandes, Saarbrücken, Germany
e-mail: santanu.pal@uni-saarland.de

S.K. Naskar
Jadavpur University, Kolkata, India
e-mail: sudip.naskar@cse.jdvu.ac.in

M.R. Costa-jussà et al. (eds.), *Hybrid Approaches to Machine Translation*,
Theory and Applications of Natural Language Processing,
DOI 10.1007/978-3-319-21311-8_3

essentially reliant on the quality of word alignment. Word alignment is not an end task in itself and is usually used as an intermediate step in SMT. Word alignment is the task of detecting correspondences between words that are translations of each other from parallel sentences. Existing statistical word alignment algorithms do not cope well with many-to-many word links and SMT Models suffer from this shortcoming of alignment algorithms to process such links.

Existing unsupervised word alignment models are based on IBM models 1–5 (Brown et al. 1993) and the HMM model (Vogel et al. 1996; Och et al. 2003). IBM Models 3, 4 and 5 are based on fertility-based models, which are asymmetric. To improve word alignment quality, the Berkeley Aligner uses the symmetric property by intersecting alignments induced in each translation direction.

In addition, in any language, Multiword Expressions (MWEs) cause major problems and they pose big challenge in statistical machine translation. MWE can be roughly defined as idiosyncratic interpretations that cross word boundaries (Sag et al. 2002). The meaning of MWEs cannot be always derived from their component words; each of which have their own separate meanings when they occur independently.

Named Entity is considered as MWEs, because it contains more than one words and used as a single semantic unit in a sentence. Named entities (NE), particularly multiword NEs, on the source and the target sides of the parallel corpus should be aligned and translated as a whole. This is also true for multiword expressions (MWE) and complex predicates in general (Pal et al. 2011). However, in the state-of-the-art PB-SMT systems, the constituents of such multiword expressions are often marked and aligned as part of consecutive phrases since PB-SMT (or any other approaches to SMT) does not generally treat multiword expressions as special tokens. This motivated us to consider NEs for special treatment in this work by converting them into single tokens that makes sure that PB-SMT also treats them as a whole.

Word alignment is one of the most difficult as well as critical tasks in SMT. Sometimes some source words, appearing in both the input as well as the training set, do not correctly get translated into the SMT output because of their mapping to NULL token or erroneous mapping during word alignment. Verb phrase translation has proven itself to be a larger challenge in SMT. 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.

In the present work, we propose improvement of word alignment quality by combining several word alignment models and tables: (1) surface-to-surface GIZA++ alignment, (2) surface-to-surface Berkeley alignment, (3) root-to-root GIZA++ alignment, (4) root-to-root Berkeley alignment and (5) rule based alignment.

The first objective of the present work is to see how single tokenization and prior alignment of NEs affect the overall MT quality. The second objective is to see whether a hybrid word alignment model combining both unsupervised and semi-supervised techniques can enhance the quality of translation in SMT.

We carried out the experiments on an English—Bengali translation task. Bengali shows high morphological richness at lexical level. Language resources in Bengali are also very scarce.

The hybrid word alignment method combines three different kinds of word alignments—Giza++ word Alignment with grow-diag-final-and (GDFA) heuristic (Koehn et al. 2003), Berkeley aligner and rule-based aligner. We have followed two different strategies to combine the three different word alignment tables: union and add additional alignment algorithm. We implemented a rule based alignment model by considering several types of chunks, which are automatically identified on the source side. Each individual source chunk is translated using a baseline PB-SMT system and validated with the target chunks on the target side. The validated source-target chunks are added in the rule based alignment table. Work has been carried out into three directions: (1) several alignment tables are combined together by taking their union; (2) extra alignment pairs are added into the alignment table which is a well-known practice in domain adaptation in SMT (Eck et al. 2004; Wu et al. 2008) and (3) the alignment table is updated through semi-supervised alignment technique. The rule based alignment table is also improved using the updated hybrid word alignment model and then we further improve the entire model during the second pass of the experiment.

The correctness of the alignments is verified by manually checking the performance of the various alignment systems. We start with the combined alignment table which is produced by the add additional alignment algorithm which is described in Sect. 3.4. Initially, we take a subset of the alignments by manually inspecting from the combined alignment table. Then we train the Berkeley supervised aligner with this labeled data. A subset of the unlabeled data from the combined alignment table is aligned with the supervised model. The output is then added as additional labeled training data for the supervised training method for the next iteration. Using this bootstrapping approach, the amount of labeled training data for the supervised aligner is gradually increased. The process is continued until there are no more unlabeled training data. In this way we establish word alignments for the entire parallel corpus. The process is carried out in a semi-supervised manner.

We carried out evaluation of the proposed model using automatic evaluation metrics and observed significant improvements over the baseline models.

The remainder of the paper is organized as follows. Section 2 discusses related work. The proposed hybrid word alignment model is described in Sect. 3. Section 4 presents 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

A multilingual filtering algorithm that generates bilingual chunk alignments from Chinese-English parallel corpus was proposed in (Zhu 2005). The algorithm has three steps. First, the most frequent bilingual chunks are extracted from the parallel

corpus. Secondly, the participating chunks for alignments are combined into a cluster and finally one English chunk is generated corresponding to a Chinese chunk by analyzing the highest co-occurrences of English chunks. Bilingual knowledge can be extracted using chunk alignment (Zhu 2005). Another method of chunk alignment with bootstrapping approach described in (Pal and Bandyopadhyay 2012); they used an SMT based model for chunk translation and then aligned the source-target chunk pairs after validating the translated chunk.

To automatically extract bilingual MWEs, a log likelihood ratio based hierarchical reducing algorithm was proposed in (Ren et al. 2009). The usefulness of these bilingual MWEs in SMT is examined by integrating bilingual MWEs into the Moses decoder (Koehn et al. 2007). They also observed the highest improvement with an additional feature that identifies whether or not a bilingual phrase contains bilingual MWEs. While in (Ma et al. 2007), the authors simplified the task of automatic word alignment as several consecutive words together correspond to a single word in the opposite language by using the word aligner itself, i.e., by bootstrapping on its output. Extracting bilingual multiword expressions and using them in statistical machine translation was first proposed by (Lambert et al. 2005). They applied their MWE extraction technique on the Verbmobil corpus and found that the integration of these bilingual MWEs into the statistical alignment improves word alignment quality as well as translation accuracy. The term: pseudo-word, a kind of multiword expression, was introduced in (Duan et al. 2010). Pseudo-word is defined as a minimal sequence of consecutive words in terms of translation. They considered these pseudo-words as a translational unit and then fed into the Chinese-to-English PB-SMT Model. The model significantly outperformed the baseline PB-SMT model in both travel domain and news domain. Bilingual lexicon construction of MWES from a French—English parallel corpus using a hybrid approach was presented in (Bouamor et al. 2012). They integrated this bilingual MWE lexicon into PB-SMT and reported improvement in translation quality. However, their algorithm works only for many to many alignments and deals with highly and weakly correlated MWES in a given sentence pair. A Maximum Entropy model based approach for English—Chinese NE alignment that significantly outperforms IBM Model4 and HMM was proposed by (Feng et al. 2004). They considered 4 features: translation score, transliteration score, source NE and target NE's co-occurrence score and the distortion score for distinguishing identical NEs in the same sentence. Capitalization cues have also been used for identifying NEs on the English side. Statistical techniques are applied to decide which portion of the target language corresponds to the specified English NE, for simultaneous NE identification and translation (Moore and Robert 2003).

To improve the learning process of unlabeled data using labeled data (Chapelle et al. 2006), semi-supervised learning method is a very useful learning technique. Researchers have begun to explore semi-supervised word alignment models that use both labeled and unlabeled data. A semi-supervised training algorithm was described in (Fraser et al. 2006), where the weighting parameters are learned from discriminative error training on labeled data, and the parameters are estimated by maximum-likelihood EM training on unlabeled data. They also used a log-linear

model, which is trained on the available labeled data to improve performance. Interpolating human alignments with automatic alignments has been proposed by (Callison-Burch et al. 2004), where the alignments of higher quality gained much higher weight than the lower quality alignments. Two separate models of standard EM algorithm, which learn separately from both labeled and unlabeled data, were developed by (Wu et al. 2006). These two models are then interpolated as a learner in the semi-supervised Ada-Boost algorithm to improve word alignment. To identify highly uncertain or most informative alignment links, active learning query strategies were applied under an unsupervised word alignment model in (Ambati et al. 2010).

Intuitively, multiword NEs 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 multiword NE are marked and aligned as parts of consecutive phrases, since PB-SMT (or any other approaches to SMT) does not generally treat multiword NEs as special tokens. This is the motivation behind considering NEs for special treatment in this work by converting them into single tokens that makes sure that PB-SMT also treats them as a whole.

Another problem with SMT systems is the erroneous word alignment. Sometimes some words are not translated in the SMT output sentence because of the mapping to NULL token or erroneous mapping during word alignment. It can often be observed that verb phrase translation poses a major challenge in SMT, particularly so for English to Indic languages. The alignments between the words inside source and target verb phrases for such language pairs are mostly found to be many-to-many.

3 Hybrid Word Alignment Model

The hybrid word alignment model is described as the combination of three word alignment models as follows:

3.1 Word Alignment Using GIZA++

GIZA++ (Och et al. 2003) is a statistical word alignment tool, which incorporates all the IBM 1-5 models. GIZA++ facilitates fast development of statistical machine translation (SMT) systems. In case of low-resource language pairs the quality of word alignments is typically quite low and it also deviates from the independence assumptions made by the generative models. Although huge amount of parallel data enables the model parameters to acquire better estimation, a large number of language pairs still lack from the unavailability of sizeable amount of parallel data. GIZA++ has some drawbacks. It allows at most one source word to be aligned with each foreign word. To resolve this issue, some techniques have already been applied, such as the following one. The parallel corpus is aligned bidirectionally; then the two

alignment tables are reconciled using different heuristics, e.g., intersection, union, and most recently grow-diagonal-final and grow-diagonal-final-and heuristics have been applied. In spite of these heuristics, the word alignment quality for low-resource language pairs still remain low and calls for further improvement. We describe our approach of improving word alignment quality in the following three subsections.

3.2 Word Alignment Using Berkley Aligner

A recent advancement in word alignment is implemented in Berkeley Aligner (Liang et al. 2006) which allows both unsupervised and supervised approach to align word from parallel corpus. We initially train the model using unsupervised technique. We make a few manual corrections to the alignment table produced by the unsupervised aligner. Then we apply this corrected alignment table as gold standard training data for the supervised aligner. The Berkeley aligner is an extension of the Cross Expectation Maximization word aligner. Berkeley aligner is a very useful word aligner because it allows for supervised training, enabling us to derive knowledge from an already aligned parallel corpus or we can use the same corpus by updating the alignments using some rule based methods. Our approach deals with the latter case. The supervised technique of Berkeley aligner helps us to align those words, which could not be aligned by our rule-based word aligner.

3.3 Rule Based Word Alignment

The proposed rule based aligner aligns named entities and chunks. Figure 1 shows the architecture of the rule-based system. For NE alignment, we first identify NEs from the source side (i.e. English) using Stanford NER. The NEs on the target side

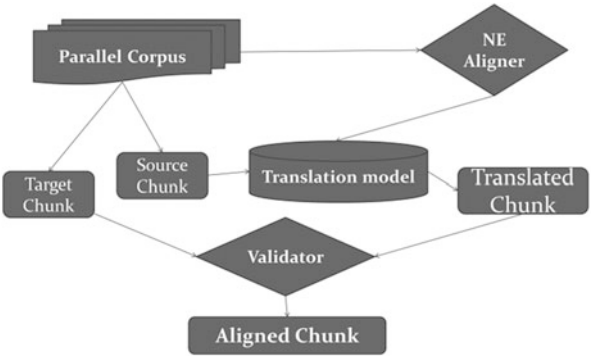

Fig. 1 System architecture of rule based aligner

(i.e. Bengali) are identified using a method described in (Ekbal and Bandyopadhyay 2009). The accuracy of the Bengali named entity recognizers (NER) is much poorer than that of English NER due to several reasons: (1) there is no capitalization cue for NEs in Bengali; (2) most of the common nouns in Bengali are frequently used as proper nouns; (3) suffixes (case markers, plural markers, emphasizers, specifiers) get attached to proper names in Bengali. Bengali shallow parser has been used to improve the performance of NE identification by considering proper names as NE. Therefore, NER and shallow parser are jointly employed to detect NEs from the Bengali sentences. The source NEs are then transliterated using a modified joint source-channel model (Ekbal et al. 2006) and aligned to their target side equivalents following the approach of (Pal et al. 2010). Since Bengali NEs differ in their choice of 'matras' (vowel modifiers), both the NEs found in the Bengali sentence as well the transliterated (i.e., Bengali) NEs are transformed into a canonical form after omitting their *matras*. The transliterated NEs are then matched with the corresponding parallel target NEs and finally we align the NEs if a match is found.

After identification of multiword NEs on both sides, we pre-processed the corpus by replacing space with the underscore character ('_'), this ensures that the multiword NEs are single tokenized and considered as a single unit. We have used underscore ('_') instead of hyphen ('-') since there already exists some hyphenated words in the corpus. The use of the underscore ('_') character also facilitates to detokenize the single-tokenized NEs after decoding.

3.3.1 Automatic Alignments of NEs Through Transliteration

We extract the source and target (single token) NEs from the NE-tagged parallel translations in which both sides contain at least one NE. Then we first create an NE parallel corpus. In the example mentioned below, we extract the NE translation pairs given in (2) from the sentence pair shown in (1), where the NEs are shown in italics.

(1a) *Kirti_Mandir*, where *Mahatma_Gandhi* was born, today houses a photo exhibition on the life and times of the *Mahatma*, a library, a prayer hall and other memorabilia.

(1b) *কির্তী_মন্দির* , যেখানে *মহাত্মা_গান্ধী* জন্মেছিলেন , বর্তমানে সেখানে *মহাত্মার* জীবন ও সেই সময়ের ঘটনাসমূহের একটি চিত্রপ্রদর্শনশালা , একটি লাইব্রেরী ও একটি প্রার্থনা ঘর এবং অন্যান্য স্মৃতিবিজড়িত জিনিসপত্র আছে ।

(2a) Kirti_Mandir Mahatma_Gandhi Mahatma

(2b) কির্তী_মন্দির মহাত্মা_গান্ধী মহাত্মার

Next, we try to align the extracted source and target NEs, as illustrated in (2). If both sides contain only one NE then the alignment is trivial, and we add such NE pairs to seed another parallel NE corpus that contains examples having only one token in both side. Otherwise, we establish alignments between the source and target NEs using transliteration. We use the joint source-channel model of transliteration (Ekbal et al. 2006) for this purpose.

If both the source and target side contains n number of NEs, and the alignments of n-1 NEs can be established through transliteration or by means of already existing alignments, then the nth alignment is trivial. Similarly, for multiword NEs, intra-NE word alignments are established through transliteration or by means of already existing alignments. For a multiword source NE, if we can align all the words inside the NE with words inside a target NE, then we assume they are translations of each other.

Since the source side NER is much more reliable than the target side NER, we transliterate the English NEs, and try to align them with the Bengali NEs. We take the 5 best transliterations produced by the transliteration system for an English word, and compare them against the Bengali words. Here, we first normalize both Bengali words: target NEs and the transliterated ones, because Bengali NEs often differ in their choice of *matras* (vowel modifiers). Thus we transform Bengali NE word into a canonical form by dropping the *matras*, and then compare the results; if they match, then we align the English NE word with the Bengali NE word.

(3) নিরজ (ন + ি + র + জ) -- নীরাজ (ন + ী + র + া + জ)

The example in (3) illustrates the procedure. Assume we are trying to align "Niraj" with "নীরাজ". The transliteration system produces "নিরজ" from the English word "Niraj" and we compare "নিরজ" with "নীরাজ". Since the consonant sequences match in both words, "নিরজ" is considered a spelling variation of "নীরাজ", and the English word "Niraj" is aligned to the Bengali word "নীরাজ".

In this way, we achieve word-level alignments, as well as NE-level alignments. (4) shows the alignments established from (1). The word-level alignments help to establish new word/NE alignments. Word and NE alignments obtained in this way are added to the parallel corpus as additional training data.

(4a) Kirti-Mandir—কির্তী-মন্দির
(4b) Kirti—কির্তী
(4c) Mandir—মন্দির
(4d) Mahatma-Gandhi—মহাত্মা-গান্ধী
(4e) Mahatma—মহাত্মা
(4f) Gandhi—গান্ধী
(4g) Mahatma—মহাত্মার

3.3.2 Automatic Chunk Alignment

For chunk alignment, the source sentences of the parallel corpus are parsed using Stanford POS tagger. The chunks of the sentences are extracted using CRF chunker. The chunker detects the boundaries of noun, verb, adjective, adverb and prepositional chunks from the sentences. In case of prepositional phrase chunks, we have taken a special attention: we have expanded the prepositional phrase chunk by examining a single noun chunk followed by a preposition or a series of noun

chunks separated by conjunctions such as 'comma', 'and' etc. For each individual chunk, the head word is identified. Similarly, target side sentences are parsed using a shallow parser. The individual target side Bengali chunks are extracted from the parsed sentences. The head words for all individual chunks on the target side are also marked. If the translated head word of a source chunk matches with the headword of a target chunk then we hypothesize that these two chunks are translations of each other.

The extracted source chunks are translated using a baseline SMT model trained on the same corpus. The translated chunks are validated against the target chunks found in the corresponding target sentence. During the validation process, if any match is found between the translated chunk and a target chunk then the source chunk is directly aligned with the original target chunk. Otherwise, the source chunk is ignored in the current iteration for any possible alignment and is considered in the next iterations.

The extracted chunks on the source side may not have a one to one correspondence with the target side chunks. The alignment validation process is focused on the proper identification of the head words and not between the translated source chunk and target chunk. The matching process has been carried out using a fuzzy matching technique. If both sides contain only one chunk after aligning the remaining chunks then the alignment is trivial. After aligning the individual chunks, we also establish word alignments between the matching words in those aligned chunks. Thus we get a sentence level source-target word alignment table.

Figure 2 shows how word alignments are established between a source-target sentence pair using the rule based method. Figure 2a shows the alignments obtained through rule-based method. The solid links are established through transliteration (for NEs) and translation. The dotted arrows are also probable candidates for intra-

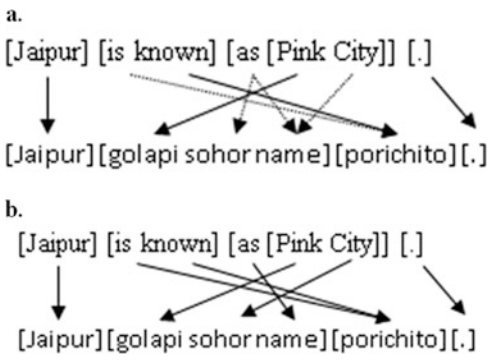

Fig. 2 Establishing alignments through rule based methods. (**a**) Rule based alignments. (**b**) Gold standard alignments

chunk word alignments; however they are not considered in the present work. Figure 2b shows the gold standard alignments for this sentence pair.

3.4 Hybrid Word Alignments Model

The hybrid word alignment method combines word alignments produced by three different kinds of word aligners—Giza++ with grow-diag-final-and (GDFA) heuristic, Berkeley aligner and rule based aligner. We have followed two different strategies to combine the three different word alignment tables.

3.4.1 Union

In the union method all the alignment tables are united together and duplicate entries are removed. Taking union of the alignments should improve the recall of the word alignment.

3.4.2 ADD Additional Alignments

In this method, we consider either of the alignments generated by GIZA++ (A_1) or Berkeley aligner (A_2) as the standard alignment as the rule-based aligner (A_3) fails to align many words in the parallel sentences. For any set of alignments {A_1, A_2, ... A_n}, we propose an alignment combination method as described in Algorithm 1.

Algorithm 1

Step 1: Choose a Standard alignment table (A_s) from the set of alignment tables {A_1, $A_2 \ldots A_n$} with the exception that any rule based alignment cannot be assigned to A_s.

Step 2: Correct the alignments in A_s using the remaining (n-1) alignment tables. Take intersection of the other n-1 alignment tables. E.g., for three alignment tables A_1, A_2 and A_3, if A_2 is assigned to A_s then find additional alignments from A_1 and A_3 using $A_1 \cap A_3$ and add these additional entries to A_s.

3.5 Berkeley Semi-Supervised Alignment

The correctness of the alignments is verified by manually checking the performance of the various alignment systems. We start with the combined alignment table, which is produced by Algorithm 1. Initially, we take a subset of the alignments, a set of

500 alignments from the combined alignment table, which was manually inspected and corrected. Then we train the Berkeley supervised aligner with this labeled data. A subset of the unlabeled data, i.e., alignments collected from the combined alignment table, is aligned with this supervised model. The output is then added as additional labeled training data for the supervised training method for the next iteration. Using this bootstrapping approach, the amount of labeled training data for the supervised aligner is gradually increased. The process is continued until there are no more unlabeled training data left. In this way we refine the whole alignment table for the entire parallel corpus. The process is carried out in a semi-supervised manner.

The manual correction process involves correction of one-to-one, one-to-many, many-to-one and many-to-many alignments. To optimize the manual effort involved we focus only on one-to-one alignment correction, other types of correction are automatically taken care of by the system during the iterative process. We manually inspected 500 alignments and observed that the quality of the one-to-one alignments is better than the other kinds of alignments. Table 1 shows statistics over the 500 manually inspected alignments.

Since the one-to-one alignment list has better accuracy, the one-to-one alignments are considered initially for correction in the 1st Iteration. In the 1st iteration of the statistical model, the manually checked 500 alignments are used with the large set of alignment. At the end of Iteration 1, it was found that the accuracy of both the one-to-one and one-to-many mapped word alignments increases as more and more words are now correctly aligned. After an in depth study of the one-to-one aligned pairs for a few word, it was found that the number of incorrectly aligned entries before the 1st iteration were more than the correctly aligned entries. A detailed analysis of the word alignment quality after 1st iteration exposed that not only this process improves the accuracy of one-to-one world alignments, the accuracy of other kinds of word alignments also improves. The example given below depicts the improvement the in word alignment.

English sentence: This variety is replicated in the food, architecture, music and culture of Brazil.

Bengali Sentence (English gloss): Brajilera khadya, parikaṭhamo, sangita, sanṣkr̥tite ei baicitra pratiphalita haya.

The example in Table 2 shows that, before the first iteration the word “replicated” is aligned to 3 Bengali words in the target side while the word “culture” remains unaligned. After the first iteration, the word “culture” is correctly

Table 1 Word alignment accuracy

Alignment	Accuracy (%)
1:1	83.2
1:2	67.4
1:3	49.1

Table 2 Word alignment improvement with iterations

Word	Alignment	
	Iteration 1	Iteration 2
NULL	7	7
This	9	9
variety	10	10
is	NA	12
replicated	8 11 12	11
in	NA	NA
the	NA	NA
food	2	2
,	NA	NA
architecture	4	4
,	5	5
music	NA	NA
and	NA	NA
culture	NA	8
of	NA	NA
Brazil	1	1
.	13	13

mapped to the target word “sangaskritite”, as these one-to-one mapped words are manually corrected in the training alignment set, the system identifies the correct alignment pairs during the successive iterations. In iteration 2, the system correctly aligns “culture” with “sangaskritite”, “is” with “hay” and “replicated” with “pratiPalita”.

For the successive iterations the correction of one-to-one mapped word alignments are preferred again. During successive iterations, the correction effort is gradually less and the accuracy of the one-to-many as well as other types of word alignment increases.

The hybrid word alignment model has been incorporated into the SMT workflow as shown in Fig. 3.

4 Tools and Resources Used

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

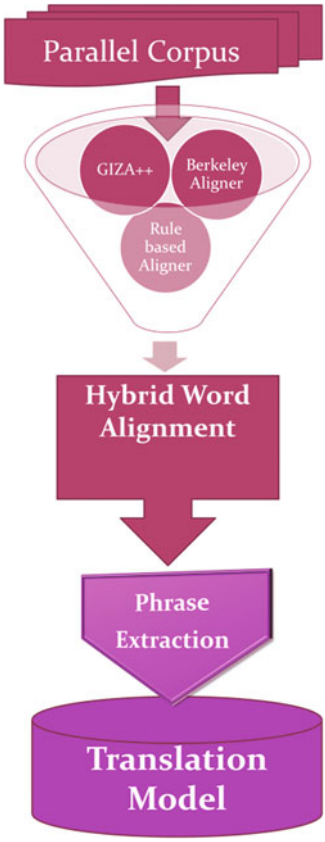

Fig. 3 Translation model using hybrid word alignment

Table 3 Corpus statistics

Corpus		# Sentence	# Words
Training	English	22,492	561,881
	Bengali	22,492	478,568
Development	English	500	10,945
	Bengali	500	9881
Monolingual	Bengali	33,597	506,859
Test	English	500	11,328
	Bengali	500	9894

English to Indian Languages Machine Translation (EILMT) System—Phase II" .[1] Table 3 presents the statistics about the dataset.

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

The Stanford Parser[2] and CRF chunker[3] have been used for identifying chunks and Stanford NER has been used to identify named entities in the source side of the parallel corpus.

The target side (Bengali) sentences are parsed by using the tools obtained from the consortium mode project "Development of Indian Language to Indian Language Machine Translation (IL-ILMT) System - Phase II[4]".

NEs in Bengali are identified using the NER system of (Ekbal and Bandyopadhyay 2009). We use the Stanford Parser, Stanford NER and the NER for Bengali along with the default model files provided, i.e., with no additional training.

The effectiveness of the present work has been tested by using the standard log-linear PB-SMT model as our baseline system: phrase-extraction heuristics described in (Koehn et al. 2003), MERT (minimum-error-rate training) (Och and Franz 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. Statistical significance tests were carried out using bootstrap resampling method (Koehn 2004) considering $p = 0.05$.

5 Experiments and Results

We randomly selected 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 506,895 words from the tourism domain was used for building 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 produced the optimum baseline result. We carried out the rest of the experiments using these settings.

We experimented with the system over various combinations of word alignment models. Our hypothesis focuses mainly on the theme that improvement in word alignment will result in improvement of the system performance in terms of translation quality, particularly so for language pairs having only small amount of training data.

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

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

[4] 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.

Table 4 Evaluation results for different experimental setups

Experiment	Exp. no.	BLEU	NIST
Baseline system using GIZA++ with GDFA	1	10.92	4.13
PB-SMT system using Berkeley Aligner	2	11.42	4.16
Experiment 1 + root-root GIZA++ alignment	3	11.08	4.14
Experiment 2 + root-root Berkeley alignment	4	11.61	4.18
Union of all alignments	5	11.22	4.15
PB-SMT system with hybrid alignment by considering (a) GIZA++ as the standard alignment (b) Berkeley alignment as the standard alignment	6a[a]	15.77	4.34
	6b[a]	16.42	4.42
Single-tokenized NE + Experiment 3	7	11.84	4.18
Single-tokenized NE + Experiment 4	8	12.02	4.20
Single-tokenized NE + (a) Experiment 6a (b) Experiment 6b	9a[a]	16.98	4.47
	9b[a]	17.72	4.53
PB-SMT system with semi-supervised Berkeley Aligner + Single-tokenized NE	10[a]	21.17	4.74

[a]Systems produce statistically significant improvements on BLEU over the baseline system

141,821 chunks were identified from the source corpus, of which 96,438 (68 %) chunks were aligned by the system. 39,931 and 28,107 NEs were identified from the source and target sides of the parallel corpus respectively, of which 22,273 NEs are unique in English and 22,010 NEs in Bengali. A total of 14,023 NEs have been aligned through transliteration.

The experiments were carried out with various experimental settings: (1) single tokenization of NEs on both sides of the parallel corpus, (2) using Berkeley Aligner with unsupervised training, (3) union of the several alignment models: rule based, GIZA++ and Berkeley Alignment, root-to-root GIZA++ alignment, root-to-root Berkeley alignment, (4) hybridization of the three alignment models, and (5) supervised Berkeley Aligner. Extrinsic evaluation was carried out on the MT quality using BLEU (Papineni et al. 2002) and NIST (Doddington 2002).

In Table 4, the baseline system (experiment 1) is the state-of-art PB-SMT system where GIZA++ with grow-diag-final-and is used as the word alignment model. Experiment 2 provides better results than experiment 1 which signifies that Berkeley Aligner performs better than GIZA++ for the English-Bengali word alignment task. Experiments 3 and 4 are carried out with root-to-root alignment; i.e. both the source and the target words are stripped to their roots and alignments are established between source and target roots, as opposed to words as is done traditionally. Root-to-root alignment helps alleviate the data sparseness problem to certain extent. It is to be noted, however, that root-to-root alignments established at the sentence level are preserved back to the word-to-word alignments. The experiments with root-to-root alignment (i.e., experiment 3 and 4) also show the same trend, i.e., Berkeley Aligner performs better than GIZA++ on root-to-root alignment. The union of all

alignments (Experiment 5) provides better scores than the baseline PB-SMT with GIZA++; however it cannot beat the results obtained with the Berkeley Aligner alone. Union of all three alignments results in improved word alignment recall; however it also introduces some noisy alignments yielding lower precision in word alignment.

In the rule based alignment table, each tuple or row provides a subset of word alignment such as NE alignment and chunk alignments in a parallel sentence. These alignments are directly incorporated into the hybrid word alignment model using Algorithm 1 (discussed in Sect. 3.4). Hybrid word alignment model with GIZA++ using root form of the source-target sentence aligned training corpus as the standard alignment (experiment 6a) and other alignments are incorporated using Algorithm 1. It produces statistically significant improvements over the baseline. Similarly, the use of Berkeley Aligner as the standard alignment of the same training data for Hybrid alignment model (experiment 6b) also results in statistically significant improvements over experiment 2 and 4. These two experiments (experiment 6a and 6b) demonstrate the effectiveness of the hybrid alignment model. It is to be noticed that the hybrid alignment model works better with the Berkeley Aligner than with GIZA++.

Single-tokenization of the NEs (experiment 7, 8, 9a and 9b) improves the system performance to some extent over the corresponding experiments without single-tokenization (experiment 3, 4, 6a and 6b); however, these improvements are not statistically significant. The Berkeley semi-supervised alignment method using a bootstrapping approach together with single-tokenization of NEs (experiment 10) provided the overall best performance in terms of both BLEU and NIST and the corresponding improvement is statistically significant on BLEU over the rest of the experiments.

6 Conclusions and Future Work

The paper proposes a hybrid word alignment model for PB-SMT. The paper also shows how effective pre-processing of NEs in the parallel corpus and direct incorporation of their alignment in the word alignment model can improve SMT system performance. In data driven approaches to MT, specifically for scarce resource language pairs, this approach can help to upgrade the state-of-the-art machine translation quality as well as the word alignment quality. The hybrid model with the use of the semi-supervised technique of the Berkeley word aligner in a bootstrapping manner, together with single tokenization of NEs, provides substantial improvements (10.25 BLEU points absolute, 93.86 % relative) over the baseline. On manual inspection of the output we found that our best system provides more accurate lexical choice as well as better word ordering than the baseline system.

As future work we would like to explore how to get the best out of multiple word alignments. We will explore other combination schemes such as majority voting

for this purpose and the concept will be tested on different sizes of training data as well as for other language pairs. Furthermore, integrating the knowledge about multiword expressions into the word alignment models is another important future direction for this work.

Acknowledgements The research leading to these results has received funding from the EU project EXPERT –the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013<tel:2007-2013>/ under REA grant agreement no. [317471].

References

Ambati, Vamshi, Stephan Vogel, and Jaime Carbonell. 2010. *10th Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing (ALNLP-2010)*, 10–17.

Bouamor, Dhouha, Nasredine Semmar, and Pierre Zweigenbeaum. 2012. Automatic construction of a multiword expressions bilingual lexicon: A statistical machine translation evaluation perspective. In *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon (CogALex-III)*, *COLING 2012*, 95–108. Mumbai.

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.

Callison-Burch, Chris, David Talbot, and Miles Osborne. 2004. Statistical machine translation with word- and sentence-aligned parallel corpora. In *Association for Computational Linguistics 2004*, 175. Morristown, NJ: Association for Computational Linguistics.

Chapelle, O., B. Schölkopf, and A. Zien, ed. 2006. *Semi-supervised learning*. Cambridge, MA: MIT.

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

Duan, Xiangyu, Min Zhang, and Haizhou Li. 2010. Pseudo-word for phrase-based machine translation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 148–156. Uppsala.

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

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

Ekbal, Asif, Sudip Kumar Naskar, and Sivaji Bandyopadhyay. 2006. A modified joint source-channel model for transliteration. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, (ACL-2006)*, 191–198. Sydney.

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

Fraser, Alexander, and Daniel Marcu. 2006. Semisupervised training for statistical word alignment. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (ACL-2006)*, 769–776. Morristown, NJ

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

Koehn, Philipp. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004)*, 388–395. Barcelona.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT-NAACL 2003: Conference Combining Human Language Technology Conference Series and The North American Chapter of the Association for Computational Linguistics Conference Series*, 48–54. Edmonton.

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 *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007): Proceedings of Demo and Poster Sessions*, 177–180. Prague.

Lambert, Patrik, and Rafael Banchs. 2005. Data inferred multiword expressions for statistical machine translation. In *Proceedings of Machine Translation Summit X*, 396–403. Phuket.

Liang, Percy, Ben Taskar, and Dan Klein. 2006. *6th Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, HLT-NAACL-2006*, 104–111.

Ma, Yanjun, Nicolas Stroppa, and Andy Way. 2007. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 304–311. Prague.

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

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

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

Pal, Santanu, and Sivaji Bandyopadhyay. 2012. Bootstrapping Chunk Alignment in Phrase-Based Statistical Machine Translation. In: *Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra), EACL-2012*, 93–100, Avignon.

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 *Proceedings of the Workshop on Multiword Expression: From Theory to Application (MWE-2010), The 23rd International Conference of Computational Linguistics (Coling 2010)*, 46–54. Beijing.

Pal, Santanu, Tanmoy Chakraborty, and Sivaji Bandyopadhyay. 2011. Handling Multiword Expressions in Phrase-Based Statistical Machine Translation. Machine Translation Summit XIII (2011), 215–224. Xiamen

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

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

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

Stolcke, Andreas. 2002. SRILM—an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, vol. 2, 901–904. Denver.

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

Wu, Hua, Haifeng Wang, and Zhanyi Liu. 2006. Boosting statistical word alignment using labeled and unlabeled data. In *Proceedings of the COLING/ACL on Main Conference Poster Sessions*, 913–920. Morristown, NJ: Association for Computational Linguistics.

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

Zhu, Xiaojin. 2005. Semi-Supervised Learning Literature Survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison. http://www.cs.wisc.edu/_jerryzhu/pub/ssl_survey.pdf