Semi-Automatic Parallel Corpora Extraction from Comparable News Corpora

Thoudam Doren Singh and Sivaji Bandyopadhyay

***Abstract*—The parallel corpus is a necessary resource in many multi/cross lingual natural language processing applications that include Machine Translation and Cross Lingual Information Retreival. Preparation of large scale parallel corpus takes time and also demands the linguistics skill. In the present work, a technique has been developed that extracts parallel corpus between Manipuri, a morphologically rich and resource constrained Indian language and English from a comparable news corpora collected from the web. A medium sized Manipuri-English bilingual lexicon and another list of Manipuri-English transliterated entities have been developed and used in the present work. Using morphological information for the agglutinative and inflective Manipuri language, the alignment quality based on similarity measure is further improved. A high level of performance is desirable since errors in sentence alignment cause further errors in systems that use the aligned text. The system has been evaluated and error analysis has also been carried out. The technique shows its effectiveness in Manipuri-English language pair and is extendable to other resource constrained, agglutinative and inflective Indian languages.**

***Index Terms*—arallel corpora, similarity measure, bilingual lexicon, morphology, named entity list.arallel corpora, similarity measure, bilingual lexicon, morphology, named entity list.P**

I. Introduction

IN the last few years , there has been a growing interest in the multilingual corpora. Preparation of large scale parallel corpora is a time consuming process and also demands the linguistics skill though parallel corpora for some of the major languages such as the English-French Canadian Hansards [1] and Europarl parallel corpus[1] [2] involving several European languages are available. There are several languages in the world for which this critical resource is yet to be developed. Sentence level alignment would be trivial if each sentence is translated into exactly one sentence. But generally, a sentence in one language may correspond to multiple sentences in the other; sometimes information content of several sentences is distributed across multiple translated sentences. Thus there are many to many alignments at the sentence level in a parallel corpus. Even in the multilingual and multicultural Indian context, the resource is not available in the required measure for several language pairs. In this view, a simple but effective semi-automatic technique has been devised to develop a parallel corpus between Manipuri, a morphologically rich and resource constrained Indian language and English. One of the major sources of such a resource is the web. The comparable news available between two languages can be collected and parallel corpora can be developed by proper filtering and processing from the raw comparable corpora. The Manipuri and English languages have been considered for case study in the present work.

Manipuri is a scheduled Indian language spoken mainly in the state of Manipur in India and in the neighboring countries namely Bangladesh and Myanmar approximately by three million people. It is a Tibeto-Burman language and highly agglutinative in nature, monosyllabic, influenced and enriched by the Indo-Aryan languages of Sanskrit origin and English. The affixes play the most important role in the structure of the language. A clear-cut demarcation between morphology and syntax is not possible in this language. In Manipuri, words are formed in three processes called affixation, derivation and compounding. The majority of the roots found in the language are bound and the affixes are the determining factor of the class of the words in the language. Annotated corpus, bilingual dictionaries, name dictionaries, WordNet, morphological analyzers, POS taggers, spell checkers etc. are not yet available in Manipuri in the required measure. Recently, manual development of sentence aligned parallel corpora in tourism domain between English and six different Indian languages, namely, Hindi, Bengali, Marathi, Oriya, Urdu and Tamil has been started under the Government of India, Department of Information Technology sponsored consortium project "Development of English to Indian Languages Machine Translation (EILMT) Systems". Manual alignment unduly constrains the volume of aligned sentences which can be retrieved given limited time and resource. There is no parallel corpus for many other Indian languages and Manipuri is one of them. In this background, an attempt has been made to extract sentence aligned parallel Manipuri-English corpora from comparable news corpora collected from the web.

The rest of the paper is organized as follows. Related works are discussed in section 2 and collection of comparable news corpora from the web is described in section 3. The preprocessing of the collected comparable news corpora and lexicon preparation are detailed in section 4. The paragraph and sentence level alignment processes are described in section

Manuscript received February 15, 2010. Manuscript accepted for publication May 31, 2010.

Authors are with Computer Science and Engineering Department, Jadavpur University, Kolkata, India (thoudam.doren@gmail.com, sivaji_cse_ju@yahoo.com).

[1]http://www.statmt.org/europarl/

5. The proposed techniques are evaluated in section 6 and the conclusion is drawn in section 7.

II. Related works

There are three kinds of sentence alignment approaches: the lexical approach, the statistical approach and the combinations of them. The performance tends to deteriorate significantly when these approaches are applied to complex corpora that are widely different from the training corpus and/or includes less literal/lexical translation. The major advantage of statistical measures is language independence. The major problem, however, is the proper selection of text units for the consideration. The chosen text units have to be comparable in their semantic complexity; otherwise statistical measures produce incorrect and incomplete results. The string similarity approach aims to extract closely related word pairs. The method is applicable to related language pairs only. Several sentence alignment techniques have been proposed that are mainly based on word correspondence, sentence length, and hybrid approaches. Word correspondence was used by Kay [3] and is based on the idea that words that are translations of each other will have similar distributions in the source (SL) and target language (TL) texts. Sentence length methods are based on the intuition that the length of a translated sentence is likely to be similar to that of the source sentence. Brown, Lai and Mercer [4] used word count as the sentence length, whereas Gale and Church [1] used character count. Brown, Lai and Mercer [4] assumed prior alignment of paragraphs. Gale and Church [1] relied on some previously aligned sentences as 'anchors'. Word correspondence was further developed in the IBM Model-1 [5] for statistical machine translation. Simard and Plamondon [6] used a composite method in which the first pass aligns at the character level as in [7] (itself based on cognate matching) and the second pass uses IBM Model-1, following Chen [8] . Composite methods are used so that different approaches can complement each other. The Gale and Church [1] algorithm is similar to the Brown [4] algorithm except that the former works at the character level while the later works at the word level. Dynamic programming is applied to search for the best alignment. It is assumed that large corpora is already subdivided into smaller chunks. News articles alignment based on Cross Lingual Information Retrieval (CLIR) are reported in [9] and [10] . Alignment of Japanese-English articles and sentences is discussed in [11] . Comparison, selection and use of sentence alignment algorithms for new language pairs are discussed in Singh [12] . Bilingual text matching using bilingual dictionary and statistics are discussed in [13] .

III. Collection of comparable news corpora from the web

The Manipuri-English comparable news corpora is collected from news available in both Manipuri and English from the website http://www.thesangaiexpress.com/ covering the period from May 2008 to November 2008 on daily basis since there is no repository maintained in the website. The corpora is comparable in nature as identical news events are discussed in both Manipuri and English news stories but these stories are not aligned either at article or sentence level. The available news covers national and international news, brief news, editorial, letter to editor, articles, sports etc. The local news coverage is more than the national and international news. The Manipuri side of the news is available in PDF format and the English side of the news is available in ASCII plain text format. A technique has been developed to convert contents from PDF documents to Unicode format. There are 15-20 common articles in each day in both the languages even though these articles are not the exact translations of each other . So, identification of the comparable articles is done from the publication of each day . From this collection, 23375 English and 22743 Manipuri sentences respectively are available in the comparable news corpus. The length of Manipuri sentences range from 10-30 words and the average length is 22.5 words per sentence. The individual articles with multiple paragraphs are reduced to single paragraphs. Use of abbreviation is very common and presence of such a list of abbreviations is necessary to improve the alignement score. The corpus cleaning process removes undesirable parts from texts such as headlines, place of news, date etc.

IV. Preprocessing of Manipuri sentences and lexicon preparation

A. Conversion from PDF to Unicode Format

The Manipuri side of the news is available in PDF format. A tool has been developed to convert Manipuri news PDF articles to Bengali Unicode[2]. The Bengali Unicode characters are used to represent Manipuri as well. The conversion of PDF format into Unicode involves the conversion to ASCII and then into Unicode using mapping tables between the ASCII characters and corresponding Bengali Unicode. The mapping tables have been prepared at different levels with separate tables for single characters and conjuncts with two or more than two characters. The single character mapping table contains 72 entries and the conjunct characters mapping table consists of 795 entries. There are conjuncts of 2, 3 and 4 characters. Sub-tables for each of the conjuncts are prepared. The preparation of such mapping table for different combination of 2,3 and 4 characters is a repetitive and time consuming process. The corpus is searched to find conjuncts with maximum number of characters (i.e., four) from the ASCII version of Manipuri file and if not found the process is repeated for conjuncts with lesser number of characters and so on. Once match is found the corresponding unicode characters are copied from the mapping table and the process is repeated for the remaining characters. English words are sometimes present in the Manipuri side of the news and these are filtered out to avoid unknown character features during the similarity-based alignment using bilingual

[2]http://unicode.org/charts/PDF/U0980.pdf

lexicon. The unknown characters are filtered and spellings are corrected manually.

B. Preparation of bilingual lexicon and parallel named entities list

The Manipuri-English lexicon [14] is being digitized and currently contains 9618 Manipuri words and the corresponding English words. Use of transliterated English words is very prominent in Manipuri. A list of 2611 Manipuri words and their English transliterations has been developed from the news corpus to improve the alignment quality. Names of people, places, and other entities often do not appear in the bilingual lexicon. The named entities which include person name, name of place and name of organisation are identified from the text based on the work of Named Entity Recognition (NER) for Manipuri using Support Vector Machine (SVM) machine learning technique [15] and transliterated using the Modified Joint Source Channel Model for Transliteration [16] . A total number of 58291 named entites have been identified in the Manipuri news side of 22743 sentences which is accountable for 11.39 % of the total words. Thus, the identification of the named entities is important and is playing a vital role in sentence aligned parallel corpora extraction for news domain.

1) Manipuri Named Entity Recognition: A part of the Manipuri news corpus of 28,629 wordforms has been manually annotated as training data with the major named entity (NE) tags, namely person name, location name, organization name and miscellaneous name to apply Support Vector Machine (SVM) based machine learning technique. Miscellaneous name includes the festival name, name of objects, name of building, date, time, measurement expression and percentage expression etc. The SVM based system makes use of the different contextual information of the words along with the variety of word-level orthographic features that are helpful in predicting the NE classes.

NE identification in Indian languages as well as in Manipuri is difficult and challenging as:

– Unlike English and most of the European languages, Manipuri lacks capitalization information, which plays a very important role in identifying NEs.
– A lot of NEs in Manipuri can appear in the dictionary with some other specific meanings.
– Manipuri is a highly inflectional language providing one of the richest and most challenging sets of linguistic and statistical features resulting in long and complex wordforms.
– Manipuri is a relatively free word order language. Thus NEs can appear in subject and object positions making the NER task more difficult compared to other languages.

The Manipuri NE tagging system includes two main phases: training and classification. The training process has been carried out by YamCha[3] toolkit, an SVM based tool for detecting classes in documents and formulating the NE tagging task as a sequence labeling problem. For classification, the TinySVM-0.07[4] classifier has been used that seems to be the best optimized among publicly available SVM toolkits.

In the present work, the NE tagset used have been further subdivided into the detailed categories in order to denote the boundaries of NEs properly. Table I shows the examples.

TABLE I
NAMED ENTITY TAGSET.

NE Tag	Meaning	NE Examples
B-LOC	Beginning, Internal or the End of a multiword location name	ইথম (Itham)
I-LOC		মোইরাং (Moirang)
E-LOC		পুরেল (Purel)
PER	Single word person name	ইরাবত (Irabot)
LOC	Single word location name	হিয়াংথাং (Hiyangthang)
ORG	Single word organization name	এআর (AR)

The best feature set (F) of Manipuri NER is identified as F=[prefixes and suffixes of length upto three characters of the current word, dynamic NE tags of the previous two words, POS tags of the previous two and next two words, digit information, length of the word].

2) Manipuri-English Transliteration: A transliteration system takes as input a character string in the source language and generates a character string in the target language as output. The process can be conceptualized as two levels of decoding: segmentation of the source string into transliteration units; and relating the source language transliteration units with units in the target language, by resolving different combinations of alignments and unit mappings. The problem of machine transliteration has been studied extensively in the paradigm of the noisy channel model. Translation of named entities is a tricky task: it involves both translation and transliteration. For example, the organization name Jadavpur viswavidyalaya is translated to Jadavpur University in which Jadavpur is transliterated to Jadavpur and viswavidyalaya is translated to University. Manipuri-English transliteration is based on Modified Joint Source Channel Model for transliteration [16].

A medium sized bilingual training corpus has been developed that contains entries mapping Manipuri names to their respective English transliterations. Transliteration units (TUs) are extracted from the Manipuri and the corresponding English names, and Manipuri TUs are associated with their English counterparts along with the TUs in context.

V. PARAGRAPH AND SENTENCE ALIGNMENT PROCESS

A. Paragraph Alignment

The Manipuri-English sentence alignment system is a two-step process. The schematic diagram of the sentence alignment process from the comparable news corpora is shown in Table I. As an initial step, the relevant articles of both sides are sorted out manually. The quantity and quality of

[3]http://chasen-org/ taku/software/yamcha/

[4]http://cl.aist-nara.ac.jp/ taku-ku/software/TinySVM

the output will decrease if less structured texts are used even if a large set of translation equivalents is used in the initial step. For highly structured texts like technical documentation, this method provides fast and precise results. An advantage is that any dictionary may be used by the algorithm as long as it suits the domain of the corpus. There were situations of many-to-one and one-to-many paragraphs between Manipuri and English articles and all the paragraphs in each articles are manually merged in one. The paragraphs are manually aligned since the boundaries are clearly marked.

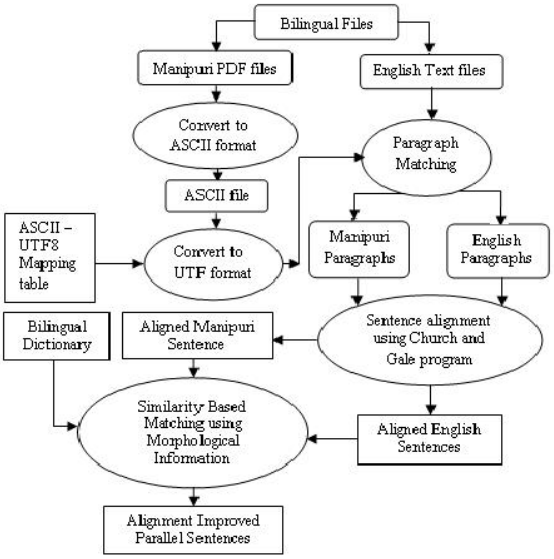

Fig. 1. Schematic diagram of sentence aligned parallel corpora extraction

After the paragraphs are aligned, the sentences are aligned using the sentence alignment program of Gale and Church [1]. However the alignment achieved at this stage is not usable mainly because the sentence alignment program is based on a simple statistical model of character lengths. It is observed that the alignment quality using this approach is poor between a highly agglutinative Indian language like Manipuri and not so agglutinative language like English.

B. Gale and Church Sentence Alignment Method

The Gale and Church program [1] uses the fact that longer sentences in one language tend to be translated into longer sentences in the other language, and that shorter sentences tend to be translated into shorter sentences. A probabilistic score is assigned to each proposed correspondence of sentences, based on the scaled difference of lengths of the two sentences (in characters) and the variance of this difference. This probabilistic score is used in a dynamic programming framework to find the maximum likelihood.

In the following, the distance function $d(x_1, y_1; x_2, y_2)$, is defined in a general way to allow insertion, deletion, substitution, etc. x and y are sequences of objects, represented as non-zero integers to be aligned. Thus let
1. $d(x_1, y_1; 0, 0)$ be the cost of substituting x_1 with y_1,
2. $d(x_1, 0; 0, 0)$ be the cost of deleting x_1,
3. $d(0, y_1; 0, 0)$ be the cost of insertion of y_1,
4. $d(x_1, y_1; x_2, 0)$ be the cost of contracting x_1 and x_2 to y_1,
5. $d(x_1, y_1; 0, y_2)$ be the cost of expanding x_1 to y_1 and y_2, and
6. $d(x_1, y_1; x_2, y_2)$ be the cost of merging x_1 and x_2 and matching with y_1 and y_2.

The recursive equation used in dynamic programming algorithm is given by equation [1]. Let s_i, $i = 1...I$, be the sentences of one language, and t_j, $j = 1...J$, be the translations of those sentences in the other language. Let d be the distance function, and $D(i, j)$ be the minimum distance between sentences $s_1, ...s_i$ and their translations $t_1, ...t_j$, under the maximum likelihood alignment. $D(i, j)$ is computed by minimizing over six cases (substitution, deletion, insertion, contraction, expansion, and merger) which, in effect, impose a set of slope constraints. That is, $D(i, j)$ is defined by the following recurrence with the initial condition $D(i, j) = 0$.

$$D(i,j) = \min \begin{cases} D(i, j-1) & +d(0, t_j; 0, 0) \\ D(i-1, j) & +d(s_i, 0; 0, 0) \\ D(i-1, j-1) & +d(s_i, t_j; 0, 0) \\ D(i-1, j-2) & +d(s_i, t_j; 0, t_{j-1}) \\ D(i-2, j-1) & +d(s_i, t_j; s_{i-1}, 0) \\ D(i-2, j-2) & +d(s_i, t_j; s_{i-1}, t_{j-1}) \end{cases} \quad (1)$$

C. Similarity-based approach to sentence alignment

The sentences in the aligned Manipuri and English paragraphs are aligned by a method based on Dynamic Programming (DP) matching. The Manipuri English sentences aligned using the Gale and Church program are realigned using the $align$ tool of Utiyama[5], 1-to-n or n-to-1 (1¡=n¡=6) alignments are taken care while aligning the sentences. In this section, the similarity measure [11] for aligning Manipuri and English sentences are discussed. Let M_i and E_i be the words in the corresponding Manipuri and English sentences for i-th alignment. The similarity between M_i and E_i is defined as in equation [2]:

$$SIM(M_i, E_i) = \frac{co(M_i \times E_i) + 1}{l(M_i) + l(E_i) - 2co(M_i \times E_i) + 1} \quad (2)$$

where

$$l(X) = \sum_{x \in X} f(x)$$

$f(x)$ is the frequency of word x in the sentences.

$$co(M_i \times E_i) = \sum_{(m,e) \in M_i \times E_i} \min(f(m), f(e))$$

$$M_i \times E_i = \{(m, e) | m \in M_i, e \in E_i\}$$

[5]http://mastarpj.nict.go.jp/ mutiyama/software.html#align

and $M_i \times E_i$ is a one-to-one correspondence between Manipuri and English words.

A measure that uses the similarity measures obtained during sentence alignments for paragraph alignment is defined in [11]. Thus $AVSIM(M, E)$ is defined as the similarity between a Manipuri article, M, and corresponding English article, E as given by equation [3].

$$AVSIM(M, E) = \frac{\sum_{k=1}^{m}(SIM(M_k, E_k))}{m} \quad (3)$$

where $(M_1, E_1), (M_2, E_2), ... (M_m, E_m)$ are the sentence alignments obtained by the method described in equation [2]. The sentence alignment measures in a correctly aligned article pair should have more similarity than the ones in an incorrectly aligned article pair. Consequently, article alignments with high $AVSIM$ are likely to be correct. The sentence alignment program aligns sentences accurately if the English sentences are literal translations of the Manipuri. However, the relation between English and Manipuri news sentences are not literal translations. Thus, the results for sentence alignments include many incorrect alignments. The sentence level similarity measure is defined in [11] as given by equation [4]

$$SntScore(Mi, Ei) = \frac{AVSIM(M, E)}{SIM(Mi, Ei)} \quad (4)$$

where $SntScore(Mi, Ei)$ is the similarity in the i-th alignment, (Mi, Ei), in the aligned articles M and E. When the correctness of two sentence alignments in the same article alignment is compared, the rank order of sentence alignments obtained by applying $SntScore$ is the same as that of SIM because they share a common $AVSIM$. However, when the correctness of two sentence alignments in different article alignments is compared, $SntScore$ prefers the sentence alignment with the more similar (high $AVSIM$) article alignment even if their SIM has the same value, while SIM cannot discriminate between the correctness of two sentence alignments if their SIM has the same value. Therefore, $SntScore$ is more appropriate than SIM if we want to compare sentence alignments in different article alignments, because, in general, an aligned sentence in a good article alignment is more reliable.

D. Incorporate morphological information

In order to improve the alignment quality between Manipuri and English, an affix adaptation module has been developed which uses the bilingual dictionary. There is no direct equivalence of the Manipuri case markers in English. So, establishing a word level similarity between Manipuri and English is more tedious if not impossible. Essentially, all morphological forms of a word and its translations have to exist in the bilingual lexicon, and every word has to appear with every possible case marker, which will require an impossibly huge amount of lexicon. In order to find the similarity between Manipuri and English based on the bilingual lexicon, the sentences of the Manipuri side are passed through the affix adaptation module and English side is searched for a corresponding match. By doing this, the number of matching words is increased thereby improving the similarity measures. The data sparseness problem can be reduced by applying similar techniques for other agglutinative and inflective languages. The affix adaptation module is developed based on the works on Manipuri Morphological analyzer [17] , Manipuri word classes and sentence type identification [18] , Morphology driven Manipuri POS tagger [19] and Manipuri-English MT system [20] . It is often observed that the number of mapping from a single Manipuri word to multiple English word is more. Whenever a dictionary is being compiled, spelling variants hamper the search for agreement between words, limiting the number of possible examples. Thus, making the right choice of English word for a Manipuri word is cumbersome.

1) Manipuri Morphology: There are free and bound roots in Manipuri. All the verb roots are bound roots. There are also a few bound noun roots, the interrogative and demonstrative pronoun roots. They cannot occur without some particle prefixed or suffixed to it. The bound root may form a compound by the addition of another root. The free roots are pure nouns, pronouns, time adverbials and some numerals. The bound roots are mostly verb roots although there are a few noun and other roots. The suffixes, which are attached to the nouns, derived nouns, to the adjectives in noun phrases including numerals, the case markers and the bound coordinators are the nominal suffixes. In Manipuri, the nominal suffixes are always attached to the numeral in a noun phrase and the noun cannot take the suffixes. Since numerals are considered as adjectives, the position occupied by the numerals in Manipuri may be regarded as adjective positions. There are a few prefixes in Manipuri. These prefixes are mostly attached to the verb roots. They can also be attached to the derived nouns and bound noun roots. There are also a few prefixes derived from the personal pronouns.

Pronominal prefix	Root	gender	number	Quantifier	Case

Fig. 2. Noun morphology

মচানুপীশিংনা*(ma-cha-nu-pi-sing-na)* **'by his/her daughters'**
মচানুপাশিংনা*(ma-cha-nu-pa-sing-na)* **'by his/her sons'**

Fig. 3. Noun morphology example

The $-ma$ "his/her" is the pronominal suffix and $-cha$ "child" is the noun root. The $-nu$ "human" is suffixed by $-pi$ to indicate a female human and $-pa$ to indicate a male human. The $-sing$ or $-khoy$ or $yaam$ can be used to indicate plurality. $-sing$ cannot be used with pronouns or proper

nouns and $-khoy$ cannot be used with nonhuman nouns. $-na$ meaning "by the" is the instrumental case marker.

In Manipuri language, the number of verbal suffixes is more than that of the nominal suffixes. New words are easily formed in Manipuri using morphological rules. Inflectional morphology is more productive than derivative morphology. There are 8 inflectional (INFL) suffixes and 23 enclitics (ENC). There are 5 derivational prefixes out of which 2 are category changing and 3 are non-category changing. There are 31 non-category changing derivational suffixes and 2 category changing suffixes. The non-category changing derivational suffixes may be divided into first level derivatives (1st LD) of 8 suffixes, second level derivatives (2nd LD) of 16 suffixes and third level derivatives (3rd LD) of 7 suffixes. Enclitics in Manipuri fall in six categories: determiners, case markers, the copula, mood markers, inclusive/exclusive and pragmatic peak markers and attitude markers. The categories are determined on the basis of position in the word (category 1 occurs before category 2, category 2 occurs before category 3 and so on). The verb morphology is more complex than the noun. Figure 2 gives the noun morphology and its example is given by Figure 3.

Figure 4 gives the verb morphology and the example is given by Figure 5.

Derivational Prefixation	Root	1st Level derivation	2nd level derivation	3rd level derivation	Inflection

Fig. 4. Verb morphology

চেক	খাই	রক	ক	নি
cek	*–khay*	*-rək*	*-kə*	*-ni*
crack	*-totally affect (1st LD)*	*-distal (2nd LD)*	*-potential (3rd LD)*	*–copula*

Fig. 5. Verb morphology example

VI. Evaluation

Methods and practical issues in evaluating alignment techniques are discussed in Langlais [21] . In the experiments in the present work, different cases considering different sizes of corpus, effect of noise of the source and target language other than AVSIM score are considered as mentioned below:

– Same size without noise
– Same size with noise
– Different size with noise

Test data of 5000 Manipuri-English parallel sentences have been manually prepared on news domain. The noise is the unrelated data from other corpus and 10 percent of the corpus size is added as the noise. The number of pairs in a one-to-n alignment is n. The evaluation parameters Recall (R), Precision (P) and F-score (F) are defined by equations [5], [6] and [7] respectively.

$$R = \frac{\#\,of\,correctly\,aligned\,sentence\,pairs}{total\,\#\,of\,sentence\,pairs\,aligned\,in\,corpus} \quad (5)$$

$$P = \frac{\#\,of\,correctly\,aligned\,sentence\,pairs}{total\,\#\,of\,aligned\,sentence\,pairs\,proposed\,by\,program} \quad (6)$$

$$F = \frac{2 \times (Precision \times Recall)}{(Precision + Recall)} \quad (7)$$

TABLE II
Same Corpus Size using [Bilingual Dictionary].

	500 sentences	1000 sentences	2000 sentences	5000 sentences
Precision	86.0	85.9	84.6	83.3
Recall	86.8	86.1	85.8	85.5
F-Score	86.3	85.9	85.1	84.3

The Table II gives the baseline result of the system in terms of precision, recall and F-score using equal number of source and target sentences (i.e., same corpus size). The system uses only Manipuri English bilingual dictionary.

TABLE III
Same Corpus Size using [Bilingual Dictionary + Transliterated Words].

	500 sentences	1000 sentences	2000 sentences	5000 sentences
Precision	97.0	96.9	95.6	93.3
Recall	96.8	97.1	95.8	93.5
F-Score	97.0	96.8	95.6	93.3

The Table III gives the result of the system using the transliterated entities in addition to the Manipuri English bilingual dictionary in terms of precision, recall and F-Score with equal number of source and target sentences (i.e., same corpus size). It is observed that there is a slight decline in the performance of the system as the corpus size increase.

TABLE IV
Same Corpus Size using [Bilingual Dictionary + Transliterated Words + Morphological Information].

	500 sentences	1000 sentences	2000 sentences	5000 sentences
Precision	98.9	98.8	98.3	95.3
Recall	97.4	96.6	96.3	94.2
F-Score	98.1	97.6	97.2	94.7

The Table IV gives the result of the system by integrating the morphological information along with the Manipuri-English bilingual dictionary and the list of transliterated Manipuri-English entities. It is observed that the system outperforms the baseline system even with increase in the corpus size. There is equal number of source and target sentences (i.e., same corpus size). The system is evaluated by putting 10 percent unrelated English sentences from other source as noise. The result of this experiment is given in

Table V. It is observed that when noise is introduced, the system performance decreases slightly.

TABLE V
NOISY CORPUS USING [BILINGUAL DICTIONARY + TRANSLITERATED WORDS + MORPHOLOGICAL INFORMATION].

	500 sentences	1000 sentences	2000 sentences	5000 sentences
Precision	95.9	94.5	93.5	92.7
Recall	94.2	93.9	93.2	92.1
F-Score	95.0	94.1	93.3	92.3

VII. CONCLUSION

The most important category for sentence alignment is one-to-one. The other alignments such as 1-to-n, n-to-1, n-to-m for $2 \leq n < 6$ and $2 \leq m < 6$ are discarded. The introduction of morphological information has further improved the alignment both for the same and different size corpus. The proposed system is evaluated considering the size, noise, transliterated entities and morphological information. The important alignment is the one-to-one with higher AVSIM score. They are shortlisted and checked for better alignment quality setting a threshold. The improvement over the baseline system after the introduction of the morphological information is observed overcoming the data sparseness on both the cases of clean as well as noisy test data. 10,350 parallel sentences have been collected in the first phase and it is planned that more parallel sentences will be collected using the technique in future. The performance of the system can be further improved by increasing the size of the bilingual dictionary including the transliterated list of named entities. The system gives a better output for highly agglutinative languages with constrained resources and can be extended to other Indian languages. This is the first attempt to extract parallel corpus of Manipuri and English from the web. The sentence aligned parallel corpora developed using this technique has been used in a Manipuri-English Statistical Machine Translation System.

REFERENCES

[1] W. A. Gale and K. W. Church, "A program for aligning sentences in bilingual corpora," in *Proceedings of 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, California, 1991, pp. 177–184.

[2] P. Koehn, "A parallel corpus for statistical machine translation," in *In MT Summit X*, 2005.

[3] M. Kay and M. Roscheisen, "Text translation alignment," in *Computational Linguistics*, 1993, pp. 121–142.

[4] P. F. Brown, J. C. Lai, and R. L. Mercer, "Aligning sentences in parallel corpora," in *Proceedings of 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, California, 1991, pp. 169–176.

[5] P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, and R. L. Mercer, "Mathematics of statistical machine translation: Parameter estimation," in *Computational Linguistics*, 1993, pp. 163–311.

[6] M. Simard and P. Plamondon, "Bilingual sentence alignment: Balancing robustness and accuracy," in *Machine Translation, 13(1)*, 1998, pp. 59–80.

[7] K. W. Church, "Char align: A program for aligning parallel texts at the character level," in *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio, 1993, pp. 1–8.

[8] S. F. Chen, "Aligning sentences in bilingual corpora using lexical information," in *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio, 1993, pp. 9–16.

[9] N. Collier, H. Hirakawa, and A. Kumano, "Machine translation vs. dictionary term translation - a comparison for english-japanese news article alignment," in *In COLING-ACL 98*, 1998, pp. 263–267.

[10] K. Matsumoto and H. Tanaka, "Automatic alignment of japanese and english newspaper articles using an mt system and a bilingual company name dictionary," in *In LREC-2002*, 2002, pp. 480–484.

[11] M. Utiyama and H. Isahara, "Reliable measures for aligning japanese-english news articles and sentences," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, Sapporo, Japan, 2003, pp. 72–79.

[12] A. K. Singh and S. Husain, "Comparison, selection and use of sentence alignment algorithms for new language pairs," in *Proceedings of the ACL-05: Association for Computational Linguistics Workshop*, Ann Arbor, USA, 2005, pp. 177–184.

[13] T. Utsuro, H. Ikeda, M. Yamane, Y. Matsumoto, and M. Nagao, "Bilingual text matching using bilingual dictionary and statistics," in *In COLING' 94*, 1994, pp. 1076–1082.

[14] S. I. Singh, "Manipuri to english dictionary." Imphal, India: S. Ibetombi Devi, 2004.

[15] T. D. Singh, N. Kishorjit, A. Ekbal, and S. Bandyopadhyay, "Named entity recognition for manipuri using support vector machine," in *In Proceedings of PACLIC 23*, Hong Kong, 2009, pp. 811–818.

[16] A. Ekbal, S. K. Naskar, and S. Bandyopadhyay, "A modified joint source-channel model for transliteration," in *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*. Sydney: Association for Computational Linguistics, 2006, pp. 191–198.

[17] T. D. Singh and S. Bandyopadhyay, "Manipuri morphological analyzer," in *In the Proceedings of the Platinum Jubilee International Conference of LSI*, Hyderabad, India, 2005.

[18] ——, "Word class and sentence type identification in manipuri morphological analyzer," in *In Proceedings of MSPIL*, Mumbai, India, 2006, pp. 11–17.

[19] ——, "Morphology driven manipuri pos tagger," in *In Proceedings of IJCNLP-08 Workshop on NLP for Less Privileged Languages*, Hyderabad, India, 2008, pp. 91–98.

[20] ——, "Manipuri-english example based machine translation system," in *International Journal of Computational Linguistics and Applications (IJCLA), ISSN 0976-0962*. Delhi, India: Bahri Publication, 2010, pp. 147–158.

[21] P. Langlais, M. Simard, and J. Veronis, "Methods and practical issues in evaluating alignment techniques," in *Proceedings of 16th International Conference on Computational Linguistics (COLING-96)*, 1996.