IJCLA VOL. 1, NO. 1-2, JAN-DEC 2010, PP. 201-216
RECEIVED 21/11/09 ACCEPTED 16/01/10 FINAL 09/03/10

Manipuri-English Example Based Machine Translation System

THOUDAM DOREN SINGH[†*], SIVAJI BANDYOPADHYAY[*]

[†]Center for Development of Advanced Computing (CDAC), India
[*]Jadavpur University, India

ABSTRACT

The development of a Manipuri to English example based machine translation system is reported. The sentence level parallel corpus is built from comparable news corpora. POS tagging, morphological analysis, NER and chunking are applied on the parallel corpus for phrase level alignment. The translation process initially looks for an exact match in the parallel example base and returns the retrieved target output. Otherwise, the maximal match source sentence is identified. For word level mismatch, the unmatched words in the input are translated from the lexicon or transliterated. Unmatched phrases are looked into the phrase level parallel example base; the target phrase translations are identified and then recombined with the retrieved output. The system is currently not handling multiple maximal matches or no match (full or partial) situation. The EBMT system has been evaluated with BLEU and NIST scores of 0.137 and 3.361 respectively, better than a baseline SMT system with the same training and test data.

Keywords: Example based machine translation, Manipuri – English, Sentence Level Parallel Corpus, Phrase Level Alignment, Evaluation

1 INTRODUCTION

Machine Translation (MT) is the process of translating text or speech units from a source language (SL) into a target language (TL) by using computers while preserving the meaning and interpretation. Various MT paradigms have so far evolved depending upon how the translation knowledge is acquired and used. The main drawback of Rule Based MT systems is that sentences in any natural language may assume a large variety of structures and hence translation requires enormous knowledge about the syntax and semantics of both the SL and TL. On the other hand, SMT techniques depend on how accurately various probabilities are measured. Realistic measurements of these probabilities can be made only if a large volume of sentence aligned parallel corpora is available. The requirement of SMT system for big parallel corpus and inability to get back the original translation used during training prompted the use of the EBMT paradigm for Manipuri-English MT system. An EBMT system stores in its example base the translation examples between the SL and TL. These examples are subsequently used as guidance for future translation tasks. In order to translate a new input sentence in SL, all matching SL sentences that match any fragment of the input SL sentence are retrieved from the example base, along with their translation in TL. These translation examples are then recombined suitably to generate the translation of the given input sentence.

Manipuri is a less privileged Tibeto-Burman language spoken by approximately three million people mainly in the state of Manipur in India as well as its neighboring states and in the countries of Myanmar and Bangladesh and is in the VIII Schedule of Indian Constitution with little resource for NLP related research and development. Some of the unique features of this language are tone, the agglutinative verb morphology and predominance of aspect than tense, lack of grammatical gender, number and person. Other features are verb final in word order, lack of numeral classifier and extensive suffix with more limited prefixation. Different word classes are formed by affixation of the respective markers. This is the first attempt to develop Manipuri-English machine translation using example based approach.

There is no parallel corpus available to develop Manipuri-English MT system at the first place. In our present work Manipuri-English news parallel corpora is being developed from web as an initial step using a semi-automatic approach. The translation methodology incorporated in our system is to search and identify for (a) complete sentence match (b) phrase level and finally (c) word levels and using entries from the lexicon after applying suffix removal/addition operations using a suffix adaptation module. The EBMT system developed so is compared with a baseline SMT system using Moses decoder. The rest of the paper is organized in such as way that related works are discussed in section 2, parallel corpus development at section 3, EMBT system methodology in section 4, evaluation in section 5 and conclusion in section 6.

2 Related Works

Aligning sentences in bilingual corpora based on a simple statistical model of character lengths using 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 is reported in [6]. Reliable measures for extracting valid news articles and sentence alignments of Japanese and English are reported in [12]. Statistical alignment tool such as GIZA++ [26] are used for words and phrase alignment of statistical machine translation systems. The EBMT system as reported by Makoto Nagao at a 1981 conference identified the three main components: matching fragments against a database of real examples, identifying the corresponding translation fragments and then recombining these translation fragments to give the target text. Researchers [25], [10] have considered EBMT to be one major and effective approach among different MT paradigms, primarily because it exploits the linguistic knowledge stored in an aligned text in a more efficient way. Example-based Machine Translation [13] makes use of past translation examples to generate the translation of a given input. [4] learn translation templates from English-Turkish translation examples. They define a template as an example translation pair where some components (e.g. word stems and morphemes) are generalized by replacing them with variables in both sentences. The use of morphemes as units allows them to represent relevant templates for Turkish. There is currently no template implementation in our EBMT system. EBMT systems are often felt to be best suited to a sublanguage approach, and an existing corpus of translations can often serve to define implicitly the sublanguage which the system can handle [25]. EBMT for highly inflected language with free order sentence constituents like Basque to English [18] are reported using morphemes for basic analysis. Hybrid Rule-Based – Example-Based MT using sub-sentential translation units are reported in [17]. There are reports on translating from poor to rich morphology languages [2], namely English to Czech and English to Hindi in Indian context [1]. Phrasal EBMT System for Translating English to Bengali is found at [27].

3 PREPARATION OF EXAMPLE BASES

Manipuri is a less computerized language and the parallel corpora, annotated corpora, dictionary and other lexical resources are generally not available. The following three example bases have been developed as part of the present work:

1. Manipuri-English Parallel corpora of 16919 sentences
2. Manipuri-English dictionary of 12229 entries which includes 2611 transliterated words
3. Manipuri-English – 57629 aligned phrases

3.1 Sentence alignment

The Manipuri-English parallel corpus is collected from news available in both Manipuri and English in a noisy form from http://www.thesangaiexpress.com/ . The corpora is comparable in nature as identical news events are described in both Manipuri and English news stories. There are 23375 English and 22743 Manipuri sentences respectively in the noisy corpus. A semiautomatic parallel corpus extraction approach is applied to align the corpora in order to make it usable for the Machine Translation system. As part of the process, the articles are aligned and dynamic programming approach [6] is applied to achieve the sentence pairs after making sure that there are equal numbers of articles on both sides. Based on the similarity measures [12], we allow 1-to-*n* or *n*-to-1 (1<=n<=6) alignments when aligning the sentences. Let M*i* and E*i* be the words of Manipuri and English sentences for *i*-th alignment. The similarity between *Mi* and *Ei* is calculated as:

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

where,

$l(X) = \sum_{x \in X} f(x)$, $f(x)$ is the frequency of x in the sentences.

$co(Mi \times Ei) = \sum(m,e) \in Mi \times Ei \ min(f(m), f(e))$

$Mi \times Ei = \{(m, e)|m \in Mi, e \in Ei \}$ *and* $Mi \times Ei$ is a one-to-one correspondence between Manipuri and English words.

A Manipuri stemmer is used in order to make use of a medium size dictionary since there is no Manipuri Wordnet available. After the parallel alignment and cleaning, there are 16919 parallel news sentences. The Manipuri-English dictionary [7] is being digitized and currently contains 9618 Manipuri words. Use of transliterated English words in Manipuri is very prominent and there are 2611 transliterated words.

3.2 Morphological Processing

In Manipuri, words are formed by 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 word class in the language. In this agglutinative language the numbers of verbal suffixes are more than that of the nominal suffixes. Works on morphological processing in Manipuri are found in [3] and [19].

Verb morphology does not indicate number, person, gender or pronominal agreement between the verb and its arguments. There are two derivational prefixes: an attributive prefix which derives adjectives from verbs and a nominalizing prefix which derives nouns from verbs.

A noun may be optionally affixed by derivational morphemes indicating gender, number and quantity. A noun may have one of the 5 semantic roles: agent, actor, patient, reciprocal/goal and theme. Actor and theme roles are not indicated morphologically, while all other semantic roles are indicated by an enclitic. Word class and sentence identification using morphological information is reported in [20].

3.3 POS Tagging and Chunking

Works on the POS tagging for Manipuri have been reported in [21] that describes Morphology Driven POS tagger of Manipuri as well as in [22] that uses Support Vector Machines (SVM) and Conditional Random Fields (CRF). The Manipuri tagset is the same as the 26 tagset defined for the Indian languages. The POS tagger with 26[1] tags using SVM methodology is identified as more viable for the present system

[1] http://shiva.iiit.ac.in/SPSAL2007/iiit_tagset_guidelines.pdf

because of its detailed 26 tags. The English sentences are POS tagged and chunked using fnTBL [14].

There is no evidence for a verb phrase constituent in Manipuri. The Manipuri verb clause consists of a verb (V) and its argument (i.e., noun phrase) this verb subcategorizes for. Given below are the phrase structure rules which derive sentences in Manipuri.

(1) S→ NP* V
NP* → NP NP NP …

Example of a Manipuri sentence is given here.

অপিকপা অমোৎপা অশোনবা অঙাংদু কপ্পী|
apikpa amotpa asonba angangdu kappi
|-----------------NP------------------| V
Small dirty weak that child is crying
‘The small, dirty, weak boy is crying’

A noun phrase may consist of a noun followed by derivational and inflectional morphology or a noun and adjectives, numerals and/or quantifiers. The order of these constituents within the noun phrase is relatively free.

(2) NP→ N (Adj*) (Num/Quant)
NP→ (Adj*) N (Num/Quant)

For example,

উচেক অচৌবদু ফজৈ |
uchek achoubadu phajei
That bird big is
beautiful.
|---------NP------------|

Grammatically, a sentence must consist of an inflected verb, which is a verb root and an inflectional suffix. An adverbial clause can be derived through the suffixation of clausal subordinators to a nominalized clause. The phrase structure rule which is used to generate adverbial clause is

(3) AdvP→ S’ CS

S’ is the sentence and CS is the clausal subordinator. It can be a locative marker দা (da) . e.g.,

ঐখোয়দা লাকপদা
eikhoida lakpada
To our place upon coming home
‘when coming to our place’

The SVM based chunker [11] is used. The training process has been carried out by YamCha[2] toolkit, an SVM based tool for detecting classes in documents and formulating the chunking task as a sequential labeling problem. For classification, we have used TinySVM-0.07[3] classifier that seems to be the best optimized among publicly available SVM toolkits. We train the system with 1,600 sentences of 35,120 words and used the model.

3.4 NER module

The NER system for Manipuri [23] is developed using Support vector machine considering the four major named entities tags, namely Person name, Location name, Organization name and Miscellaneous name. The training process has been carried out by YamCha toolkit, an SVM based tool for detecting classes in documents and formulating the NER tagging task as a sequential labeling problem trained with 28,629 sentences. For classification, we have used TinySVM-0.07 classifier that seems to be the best optimized among publicly available SVM toolkits. Experimental results show the effectiveness of the proposed approach with the overall average Recall, Precision and F-Score values of 93.91%, 95.32% and 94.59% respectively. The named entities are transliterated into the target language using modified joint source channel model for transliteration [28].

3.5 Word and Chunk alignment

Each Manipuri word has no one-to-one correspondence with the words of English sentences and also there is no direct equivalence of Manipuri case markers to English. Words and phrases are aligned using GIZA++, a statistical word alignment toolkit [26]. The high quality aligned phrases are extracted in order to feed into the generation module of the system. A word in Manipuri can correspond to several English words and vice versa. Some of the examples are:

ৱাথোক লান্থোক (*wathok lanthok*)←→crisis

লৌথবা (*louthaba*)←→take something down

লৌথোক লৌশিন (*louthok lousin*)←→ give and take

চাইখায়বা (*chaikhayba*)←→ scatter

[2] http://chasen.org/~taku/software/yamcha/

[3] http://chasen.org/~taku/software/TinySVM/

Also some Manipuri to English translation variations with additional suffixes but maintaining the same meaning is observed as given below:

চেৎহনবা (*chet-han-ba)* / চেৎশিনহনবা (*chet-sin-han-ba)* ←→ tighten
ঙহাক (*Ngahak)* / ঙহাক্তং (*Ngahak-tang)* ←→ a while

The variations of the verb part are caused by the inclusion/exclusion of derivational suffixes. The verbal suffixes are used to indicate the mood, aspect and not only indicating the type of sentences.

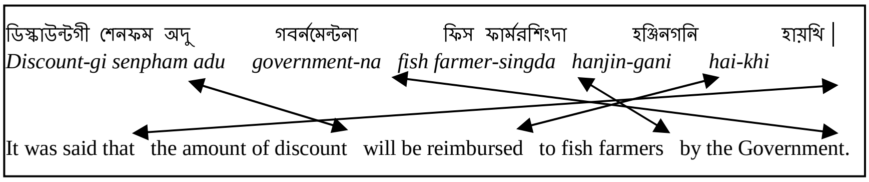

Figure 1: Equivalence between Manipuri and English components

Chunks are aligned using a dynamic programming "edit-distance style" alignment algorithm. In the following, *a* denotes an alignment between a target sequence *e* and a source sequence *f*, with $I = |e|$ and $J = |f|$. Given two sequences of chunks, we are looking for the most likely alignment $\hat{a}$:

$$\hat{a} = \underset{a}{\mathrm{argmax}}\ \mathrm{P}(a|e,f) = \underset{a}{\mathrm{argmax}}\ \mathrm{P}(a,e|f).$$

Considering alignments such as those obtained by an edit-distance algorithm, i.e.

$$a = (t_1, s_1)(t_2, s_2)\ \ldots\ (t_n, s_n),$$

with $\forall\ k \in [1, n],\ t_k \in [0, I]$ and $s_k \in [0, J]$, and $\forall\ k < k'$:

$$t_k \leq t_{k'} \text{ or } t_{k'} = 0,$$

$$s_k \leq s_{k'} \text{ or } s_{k'} = 0,$$

$$I \subseteq \bigcup_{k=1}^{n}\{t_k\},\ J \subseteq \bigcup_{k=1}^{n}\{s_k\},$$

where $t_k = 0$ and $s_k = 0$ denote a non-aligned target and source chunks. We then assume the following model:

$$\mathrm{P}(a,e|f) = \prod_k \mathrm{P}(t_k, s_k, e|f) = \prod_k \mathrm{P}(e_{t_k} | f_{s_k}),$$

where $P(e_0|f_j)$ and $P(e_i|f_0))$ denote an “insertion” and “deletion” probabilities respectively.

Assuming that the parameters $P(et_k\ |fs_k\ )$ are known, the most likely alignment is computed by a simple dynamic-programming algorithm which is a classical edit-distance algorithm in which distances are replaced by inverse-log-conditional probabilities. Moreover, this algorithm can be simply adapted to allow for block movements, in the context of MT evaluation [8]. This adaptation is necessary to take into account the potential differences between the order of constituents in Manipuri and English. We compute these parameters by relying on the information contained within the chunks considering word to word probabilities and chunk labels. Relationships between chunks are then computed using the model:

$$P(e_i|f_j)=\sum_{a_c} P(a_c, e_i|f_j) \simeq \max_{a_c} P(a_c, ei|fj)= \prod_k \max_{l} P(e_{il}\ |f_{jk}\ ).$$

In the case of chunk labels, a simple matching algorithm is used. It is possible to combine several sources of knowledge in a log-linear framework, in the following manner:

$$log P(e_i|f_j) = \sum \lambda_k log P_k(e_i|f_j) - log Z,$$

where $Pk(.)$ represents a given source of knowledge, λ_k the associated weight parameter and Z a normalization parameter. To produce a higher quality, the aligned phrases generated using GIZA++ are also added to the aligned chunks extracted by the chunk alignment module.

4 MT SYSTEM DEVELOPMENT METHODOLOGY

This is the first attempt to build MT system for Manipuri to English. While the EBMT employ pattern matching technique to translate subparts of the given input sentence , two fundamental problems of developing Manipuri to English EBMT system are (a) wide syntactic divergence between the source and target languages (b) higher degree of agglutination and richer morphology of Manipuri compared to English. Considering the first problem, we resolve it by adapting the following approach of reordering the input Manipuri sentence. Manipuri follows verb final in word order and there is lack of grammatical relation between subject and object. For example, the following sentence pair follows the same meaning (Tomba drives the car), though with different emphasis.

তোম্ব-না	কার-দু	থৌই
Tomba-na	*Car-du*	*thou-i*
Tomba-nom	Car-distal	drive

কার-দু	তোম্ব-না	থৌই
Car-du	*Tomba-na*	*thou-i*
Car-distal	Tomba-nom	drive

The identification of subject and object in both the sentences are done by the suffixes না (*na*) and দু (*du*). The case markers are the critical part of conveying right meaning during translation though the most acceptable order is SOV. The basic difference of phrase order compared to English is handled by reordering the input sentence following the rule [16]:

$C'_mS'_mS'O'_mO'V'_mV' \rightarrow SS_mV\ V_mOO_mC_m$

where, S: Subject
O: Object
V : Verb
C_m: Clause modifier
X': Corresponding constituent in Manipuri, where X is S, O, or V
X_m: modifier of X

The phrase reordering program is written using the perl module Parse::RecDescent.

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 parallel example bases, and every word has to appear with every possible case marker, which will require an impossibly huge amount of example base. Dealing at sub-sentence level replicates more complexity even at the level of chunking, before the actual process kicks off. One major advantage of EBMT is that it requires neither a huge parallel corpus as required by SMT, nor it requires framing a large rule base required by RBMT. Study of EBMT is therefore feasible for us as we do not have access to such linguistics resources. The translation steps incorporated in our system is to search and identify for (a) complete sentence match (b) phrase level and finally (c) word levels and using entries from the lexicon after applying suffix removal/addition operations using a suffix adaptation module of the source language input sentence, translate the corresponding units individually to the target language and finally arranging the translated phrases to form the target language equivalent of the source language sentence. While relating the Manipuri and English noun phrases (NPs), NPs usually end with a case-morpheme that contains information about case and number. For example, স্কুলদুগী নুপামচাশিংদুনা (*school-du-gi nupamacha-sing-du-na*, "by the boys of the school") is related as follows:

স্কুল *(school,* school*)* + দু*(du- distal marker,* the*)* + গী*(gi- case marker,* of*)* নুপামচা*(nupamacha,* boy*)*+শিং*(sing - plural marker,* s*)*+ দু*(du- distal marker,* the*)*+না*(na –case marker,* by*)*

The input sentence is passed through a stemmer in order to separate the significant suffixes along with the corresponding information for the phrase level and word adaptations. Basic sentence types in Manipuri are determined through illocutionary mood markers, all of which are verbal inflectional suffixes, with the exception of the interrogative which is an enclitic.

The simple aspect markers are -ই -y, -মি -mi, -নি -ni, -পি -pi, -ঙি -ngi, -লি -li. The progressive aspect makers are -রি –ri, -লি -li . The perfect aspect markers are -রে –re, -লে -le. To handle the various surface words of the input text, a stemmer is plugged in to maximize the matches. In the matching module, there is establishment of correspondence between units in a bilingual text at sentence, phrase or word level. Sentences can, however, be quite long. And the longer they are, the less possible it is that they will have an exact match in the translation archive, and the less flexible the EBMT system will be. In practice, EBMT systems that operate at sub-sentence level involve the dynamic derivation of the optimum length of segments of the input sentence by analyzing the available parallel corpora [5]. This requires a procedure for determining the best "cover" of an input text by segments of sentences contained in the database. It is assumed that the translation of the segments of the database that cover the input sentence is known. What is needed, therefore, is a procedure for aligning parallel texts at sub-sentence level. If sub-sentence alignment is available, the approach is fully automated but is quite vulnerable to the problem of low quality as mentioned above, as well as to ambiguity problems when the produced segments are rather small. The problem of multiple phrase matches will be handled later using the language model of the target language by picking up proper target phrase. The other alternative that will be experimented in future will be to look for most probable maximal match using frequency information for each parallel pair. If there is no match, either partial or full, in the example base, the future plan is to go for phrasal EBMT system. The algorithmic steps followed are depicted below:

a. If there is Sentence level match
    - Produce exact output translation

b. Else process the input – POS, Morph, NER and Chunks
   For maximal match (find the sentence in the Example Base that matches most with the input)
    i. one maximal match
        - phrase level mismatch - look for phrase level match and return output, replace this translated phrase in the retrieved target for the maximal match sentence as the parallel sentence level Example Base is phrase aligned
        - word level mismatch - look into the bilingual lexicon or transliteration
        - above is applicable for more than one word or phrase mismatch

    ii. more than maximal match
        - carry out the above process for all the maximal match pairs. The best target among multiple outputs is selected using the language model.
        - take the pair that occurs most in the Example Base – keep frequency information for each pair, then do as in one maximal match.

    iii. no match in the sentence level and maximal
        - go for phrasal EBMT

Finally, the translated fragments obtained so are stitched together to form the target sentence following the reordering rules as per the target language.

5 EVALUATION

The EBMT system is developed with parallel 15319 sentences, 57629 phrases and 12229 words and evaluated with 900 gold standard test sentences. We use BLEU and NIST scores for the evaluation of our system. A higher BLEU score indicates better translation. We develop a Manipuri-English baseline SMT system with the same example base used for EBMT and compare the result with EBMT system developed as shown in table 3. There is no previous report available of Manipuri-English SMT system either. The Moses [9] decoder is used. The English (trigram) language model is trained on the English portion of the training data, using the SRI Language Modeling Toolkit [24] with modified Kneser-Ney smoothing.

The two experiments of EBMT and SMT are done using 15319 sentences plus 12229 words. The testing is done three fold taking 300 sentences each.

Table 1 : Statistics of corpus used

	#sentences	#words
Parallel corpus	15319	366728
Test corpus	(300+300+300)=900	20190

Table 2: Evaluation result

Technique	Test#1		Test #2		Test#3		Average	
	BLEU	NIST	BLEU	NIST	BLEU	NIST	BLEU	NIST
Baseline SMT	0.134	3.405	0.125	3.12	0.126	3.06	**0.128**	**3.195**
EBMT system	0.150	3.513	0.131	3.25	0.132	3.32	**0.137**	**3.361**

6 CONCLUSION

The result of initial experiment of Manipuri-English EBMT system is quite encouraging with NIST score of 3.361 and BLEU score of 0.137 which is better than a baseline SMT system. Since, the source side of the translation is highly agglutinative and morphologically rich, incorporating the morphological information could help improving the system. However, the performance of the overall system can be improved further with the addition of other modules such as word sense disambiguation, multiword expression etc. By proper handling of divergence and adaptation of Manipuri-English EBMT performance can be further improved.

REFERENCES

1. Ananthakrishnan, R., Choudhary, H., Ghosh, A., Bhattacharyya, P.: Case markers and Morphology: Addressing the crux of the fluency problem in English-Hindi SMT, Proceedings of IJCNLP, Singapore, (2009)
2. Avramidis, E., and Koehn, P.: Enriching Morphologically Poor Languages for Statistical Machine Translation, Proceedings of ACL-08, HLT, (2008)
3. Choudhury, S. I., Singh, L. S., Borgohain, S., Das, P.K.: Morphological Analyzer for Manipuri: Design and Implementation, In proceedings of AACC, Kathmandu, Nepal, pp 123-129. (2004)
4. Cicekli, I., and G¨uvenir, H. A.: Learning translation templates from bilingual translation examples. In M. Carl and A. Way, editors, Recent Advances in Example-Based Machine Translation, pages 255–286. Kluwer Academic Publishers, Dordrecht, The Netherlands. (2003)
5. Cranias, L., Papageorgiou, H. and Piperidis, S.: ‘A Matching Technique in Example-Based Machine Translation’, In proceedings of Coling (1994), pages. 100–104. (2004)
6. Gale, W. A., Church, K. W.: A program for aligning sentences in bilingual corpora. Computational Linguistics, 19(1):75–102. (1993)
7. Imoba, S.: Manipuri to English Dictionary. Published by:- S. Ibetombi Devi, Imphal (2004)
8. Leusch, G., Ueffing, N., and Ney, H.: CDER: Efficient MT evaluation using block movements. In proceedings of EACL-06, pages 241-248 (2006)
9. Koehn, P., Hieu, H., Alexandra, B., Chris, C., Marcello, F., Nicola, B., Brooke, C., Wade, S., Christine, M., Richard, Z., Chris, D., Ondrej, B., Alexandra, C., Evan, H.: Moses: Open Source Toolkit for Statistical Machine Translation, Proceedings of the ACL 2007 Demo and Poster Sessions, pages 177–180, Prague. (2007)
10. Kit, C., Pan, H. and Webster, J.: Example-Based Machine Translation: A New Paradigm, Translation and Information Technology, Chinese U of HK Press, pages. 57-78. (2002)
11. Kudo, T., and Matsumoto, Y.: Use of Support Vector Learning for Chunk Identification, In Proceedings of CoNLL-2000. (2000)
12. Utiyama, M., and Isahara, H.: Reliable Measures for Aligning Japanese-English News Articles and Sentences, Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1, Pages: 72 – 79, Sapporo, Japan. (2003)
13. Nagao, M.: A framework of a mechanical translation between Japanese and English by analogy principle. In Proceedings of the International NATO Symposium on Artificial and Human Intelligence, pages 173–180, Lyon, France. (1984)
14. Ngai, G. and Florian, R.: Transformation-based Learning in the Fast Lane, Proceedings of NAACL. (2001)
15. Papineni, K., Roukos, S., Ward, T., and Zhu, W.: BLEU: a Method for Automatic Evaluation of Machine Translation, IBM Research Report, Thomas J. Watson Research Center (2001)

16. Rao, D., Mohanraj K., Hegde, J., Mehta, V. and Mahadane, P.: A Practical Framework for Syntactic Transfer of Compound-Complex Sentences for English-Hindi Machine Translation, Proceedings of KBCS 2000. (2000)
17. Sánchez-Martínez, F., Forcada, M. L., and Way, A.: Hybrid Rule-Based - Example-Based MT: Feeding Apertium with Sub-sentential Translation Units, In Proceedings of the 3rd International Workshop on Example-Based Machine Translation, pages 11-18, Dublin, Ireland. (2009)
18. Stroppa, N., Groves, D., Way, A., Sarasola, K.: Example-Based Machine Translation of the Basque Language, In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, pages 232-241, Cambridge, Massachusetts, USA (2006)
19. Singh, T. D., Bandyopadhyay, S.: Manipuri Morphological Analyzer, In the Proceedings of the Platinum Jubilee International Conference of LSI; December 6-8, 2005, University of Hyderabad, India. (2005)
20. Singh, T. D., Bandyopadhyay, S.: Word Class and Sentence Type Identification in Manipuri Morphological Analyzer. In Proceedings of MSPIL, IIT Bombay, pages 11-17. (2006)
21. Singh, T. D., Bandyopadhyay, S.: Morphology Driven Manipuri POS Tagger, IJCNLP-08 Workshop on NLP for Less Privileged Languages. Proceedings of the Workshop. AFNLP, pages 91-98, 11. January, 2008, IIIT, Hyderabad, India. (2008)
22. Singh, T. D., Ekbal, A., Bandyopadhyay, S.: Manipuri POS Tagging using CRF and SVM: A Language Independent Approach, In the proceedings of ICON-2008, Pune, India, pages 240-245 (2008)
23. Singh, T. D., Kishorjit, N., Ekbal, A,. Bandyopadhyay, S.: Named Entity Recognition for Manipuri using Support Vector Machine, In proceedings of PACLIC 23, Hong Kong (2009)
24. Stolcke, A.: SRILM – An extensible language modeling toolkit. In Proceedings of the International Conference on Spoken Language Processing, pages 901–904, Denver, Colorado. (2002)
25. Somers, H.: Review article: Example-Based Machine Translation, Machine Translation 14, pages 113-158. (1999)
26. Och, F., Ney, H.: A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, 29(1):19–51. (2003)
27. Naskar, S. K., Bandyopadhyay, S.: A Phrasal EBMT System for Translating English to Bengali, In the Proceedings of MT SUMMIT X, Phuket, Thailand. (2005)
28. Ekbal, A., Naskar, S.K., Bandyopadhyay, S.: A Modified Joint Source-Channel Model for Transliteration, Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 191–198, Sydney (2006)

THOUDAM DOREN SINGH
CENTER FOR DEVELOPMENT OF ADVANCED COMPUTING (CDAC)
GULMOHAR CROSS ROAD NO 9, JUHU, MUMBAI-400049, INDIA
E-MAIL: <THOUDAMDS@CDACMUMBAI.IN>

SIVAJI BANDYOPADHYAY
COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
JADAVPUR UNIVERSITY, KOLKATA-700032, INDIA
E-MAIL: <SIVAJI_CSE_JU@YAHOO.COM>