Integration of Reduplicated Multiword Expressions and Named Entities in a Phrase Based Statistical Machine Translation System

Thoudam Doren Singh[†]
Department of Computer Science and Engineering
Jadavpur University
Kolkata-700032, India
thoudam.doren@gmail.com

Sivaji Bandyopadhyay
Department of Computer Science and Engineering
Jadavpur University
Kolkata-700032, India
sivaji_cse_ju@yahoo.com

Abstract

The language specific Multiword expressions (MWEs) play important roles in many natural language processing (NLP) tasks. Integrating reduplicated multiword expressions (RMWEs) into the Phrase Based Statistical Machine Translation (PBSMT) to improve translation quality is reported in the present work between Manipuri, a highly agglutinative Tibeto-Burman language and English. In addition, Multiword Named Entities (MNEs) coupled with Transliterated non-named entities (non-NE) between Manipuri and English phrase based SMT system are also integrated. The tighter integration of RMWEs and NEs into the PBSMT is carried out after automatic extraction using SVM based machine learning technique followed by automatic bilingual RMWE and MNE extraction using GIZA++ alignment. Our experimental results show improvement in the PBSMT system BLEU and NIST scores over the baseline system. Subjective evaluation indicates the improvement in the adequacy.

1 Introduction

Multiword expressions (MWEs) are a key challenge for the development of large-scale, linguistically sound natural language processing technology (Sag et al, 2002). The various kinds of multiword expressions should be analyzed in distinct ways. An adequate comprehensive analysis of multiword expressions must employ both symbolic and statistical techniques. MWEs span a range of constructions, from completely frozen, semantically opaque idiomatic expressions, to frequent but morphologically productive and semantically compositional collocations. Various linguistic processes (orthographic, morphological, syntactic, semantic and cognitive) apply to MWEs in idiosyncratic ways. Notably, MWEs blur the distinction between the lexicon and the grammar, since they often have some properties of words and some properties of phrases. The MWE identification works concentrate on compound nouns, noun-verb combination, idioms and phrases for several languages such as Hindi and Hebrew but not much on RMWEs. The reason may be that the reduplicated words are either rare or easy to identify for these languages since only complete duplication and some amount of partial reduplication may be present in these languages.

On the other hand, reduplicated MWEs (RMWE) of several varieties are widely present in Manipuri[1]. In the present work, the identification of Manipuri Named Entities (NEs) and the identification and classification of RMWEs is carried out using the SVM based machine learning approach.

† Presently at Center for Development of Advanced Computing (CDAC), Mumbai, India

[1] Manipuri, locally known as Meiteilon or Meeteilon, is a less privileged, morphologically rich, highly agglutinative language spoken basically in the states of Manipur, Assam, Tripura and Mizoram in India and in the neighboring countries of Myanmar and Bangladesh approximately by three million speakers. Manipuri became the first Tibeto-Burman (TB) language to receive recognition in the year 1992 as a schedule VIII language of India.

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 1304–1312,
Chiang Mai, Thailand, November 8 – 13, 2011.

Manipuri is very rich in RMWEs like other Tibeto-Burman languages. The work of (Singh, 2000) describes the linguistic rules for identifying reduplicated words. Manipuri is the direct descendant of the conglomeration of Tibeto-Burman dialects of seven different clans. This is reflected in the commonly found use of double synonyms in Manipuri, technically known as 'semantic reduplication'. The process of reduplication (Singh, 2000) is defined as: 'reduplication is that repetition, the result of which constitutes a unit word'. These single unit words are the MWEs. The RMWEs in Manipuri are classified as: (i) Complete RMWEs, (ii) Partial RMWEs, (iii) Echo RMWEs and (iv) Mimic RMWEs. Apart from these four types of RMWEs, there are also cases of (a) Double RMWEs and (b) Semantic RMWEs.

Complete RMWEs: In the complete RMWEs the single word or phrase is repeated once forming a single unit regardless of the phonological or morphological variations. Interestingly in Manipuri these complete reduplicated MWEs can occur as Noun, Adjective, Adverb, *Wh*- question type, Verbs, Command and Request.

মরিক মরিক ('*marik marik*') which means 'drop by drop'. [Noun]

অটেক অটেকপা ('*atek atek-pa*') which means 'fresh' [Adjective]

করি করি ('*kari kari*') means 'what/which'. [Wh-question]

Partial RMWEs*:* In case of partial reduplication the second word carries some part of the first word as an affix to the second word, either as a suffix or a prefix.

For example, চত্থোক চত্সিন ('chat-thok chat-sin') means 'to go to and fro'; শামী লানমী ('*saa-mi laan-mi*') means 'army'.

Mimic RMWEs: In the mimic reduplication the words are complete reduplications but the morphemes are onomatopoetic, usually emotional or natural sounds. For example, করক করক ('*krak krak*') means 'cracking sound of earth in drought'.

Echo RMWEs: The second word does not have a dictionary meaning and is basically an echo word of the first word. For example, থকসি খাসি ('*thak-si kha-si*') means 'good manner'.

Double RMWEs: Such type of reduplication generally consists of three words where the prefix or suffix of the first two words is reduplicated but in the third word the prefix or suffix is absent. An example of double prefix reduplication is ইমুন ইমুন মুনবা ('*i-mun i-mun mun-ba*') which means, 'completely ripe'.

Semantic RMWEs: Both the reduplication words have the same meaning that is shared by the MWE itself. Such types of MWEs are very special to the Manipuri language. For example, পামবা কৈ ('paamba kei') means 'tiger' and each of the component words means 'tiger'.

The performance of the SMT is heavily dependent on the quality of the Parallel Corpora and the alignment models used. Integrating MWEs into the Machine Translation (MT) systems in general and phrase based Statistical Machine Translation (PBSMT) system in particular is a critical problem. Thus, there is the need for identifying Manipuri RMWE and integrating into the PBSMT system. In the present work, we integrate Manipuri Named entities (both single word and multiword), RMWEs and non-NE transliterated entities into the existing phrase-based model.

2 Related Works

Koehn and Knight (2003) discussed empirical methods for compound splitting by learning rules from monolingual and parallel corpora. Lambert and Banchs (2005) proposed technique for extracting bilingual MWEs based on grouping as units before performing the statistical alignment. (Ren et al., 2009) presented the Log Likelihood Ratio based hierarchical reducing algorithm to automatically extract bilingual MWE and investigated the performance of three different strategies in applying bilingual MWEs for SMT system using Moses (Koehn et al., 2007). Carpuat and Diab (2010) explored static integration strategy that segments training and test sentences according to the MWE vocabulary, and dynamic integration strategy that adds a new MWE-based feature in SMT translation lexicon. Handling of

named entities and compound verbs in PBSMT (Pal et al., 2010) has been reported in the English-Bengali task in the Indian language context. They established prior NE alignments in the parallel corpora by transliterating source NEs into the target language using modified joint source channel transliteration technique (Ekbal et al., 2006) which incorporates different contextual information into the model. In the process, the identified NEs and compound verbs are converted into a single token by replacing spaces between the constituent words by underscores.

3 Manipuri-English Parallel Corpora

The Manipuri-English Parallel corpus (Singh and Bandyopadhyay, 2010b) on News Domain is used for training and testing. The Manipuri news corpus is collected from the website http://www.thesangaiexpress.com/. On analyzing the corpus, the statistics shows that Named entities (both single word and multiword), RMWEs and non-NE transliterated entities are present in significant numbers.

The presence of NEs in the Manipuri News corpus is approximated at 11.39%. However, the presence of non-NE transliterated entities is lesser and estimated at 9.2%. The following example (a) demonstrates the usage of the transliterated non-NE in the Manipuri news.

The following translation examples demonstrate the usage of NEs, transliterated non-NEs and RMWEs in the Manipuri source sentences.

(a)

হায় পৱার কম্মিটিগী ইন্সত্রক্সন মতুং ইন্না য়ুঠ এফিয়ার্স এণ্ড স্পোর্টসকি জোইন্ট সেক্রেটরিনা চহি অসিগী এপ্রিল ২৪ দা ডিপার্টমেন্টশিংদগী অহাঙবা পোষ্টশিং পীখল্লক্লবা খঙহনখ্রবনি

According to the instruction of the high power committee, the joint secretary of youth affairs and sports notified to submit the vacant posts in the departments on 24 April of this year.

Considering the above translation example (a) between Manipuri and English sentence, the Manipuri transliteration of the non-NE is shown in Table 1. This significant presence of these entities plays an important role in the news corpus. These are the examples of enrichment of Manipuri languages using Manipuri transliteration from English.

Manipuri Transliteration	**English Words**
ইন্সত্রক্সন	Instruction
জোইন্ট	Joint
সেক্রেটরি	Secretary
এপ্রিল	April
ডিপার্টমেন্ট	Department
পোষ্ট	Post
হায়	High
পৱার	Power
কম্মিটি	Committee
য়ুঠ	Youth
এফিয়ার্স	Affairs
এণ্ড	And
স্পোর্টস	Sports

Table 1: Sample Manipuri Transliterated non-NE

(b)

মোনিকা ইম্ফালদা থোরক্লবা ক্লাইটকি টিকেত্ তান্নবদা ফংখিদবদগী থোরকপা ঙমখিদবনি

Monika could not come due to the unavailability of the flight ticket to Imphal.

From the example (b), the NEs are given in Table 2.

Manipuri Named Entities	**English Transliteration**
মোনিকা	Monika
ইম্ফাল	Imphal

Table 2: Sample Manipuri Named Entities

(c)

অশোক অপন , মীশি মীনা , অফা অপুন য়াওরক্লবদি মদুদা থোক্লকপা অফ ফত্ত পূম্নমকি দায়ত্ব পল্ফনা

পুগদবনি হায়নসু ঙসিগী মীফমনা মত অমত্তা ওইনা য়ানথি হায়রি

It is said that the meeting unanimously agreed that PULF should be held sole responsible of dire consequences in case of untoward incidents, killing and kidnapping.

From the translation example (c), the Manipuri RMWEs are given in Table 3.

Manipuri RMWE	**English Meaning**
অশোক অপন	untoward incidents
মীশি মীনা	killing
অফা অপুন	kidnapping
অফ ফত্ত	dire consequences

Table 3: Sample Manipuri RMWEs

4 NE 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 to 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*.

A bilingual training set of Manipuri NEs and their respective English transliterations, has been created by collecting Manipuri person, location and organization names from Manipuri news corpus and then manually entering their English transliterations. This bilingual training set is automatically analyzed to acquire mapping knowledge in order to transliterate new Manipuri person, location and organization names to English. Transliteration units (TUs) are extracted from the Manipuri-English name pairs and Manipuri TUs are associated with their English counterparts. Some examples are given below:

(a) মনিপুর(*manipur*) → ম । নি । পু । র

manipur → ma | ni | pu | r

(b) রাজকুমার(*rajkumar*) →রা । জ । কু । মা । র

rajkumar → ra | j | ku | ma | r

(c) অভিনন্দন(*abhinandan*)→অ । ভি । ন । ন্দ ।ন

abhinandan → a | bhi | na | nda | n

The TUs are the lexical units for machine transliteration. The Manipuri NE is divided into Transliteration Units (TU) with patterns C+M?, where C represents a consonant or a vowel or a conjunct and M represents the vowel modifier or matra. An English NE is divided into TUs with patterns C*V*, where C represents a consonant and V represents a vowel. The system automatically learns mappings from the bilingual training set of 20,000 NEs. Aligned TUs along with their contexts are automatically derived from this bilingual training set to generate the collocation statistics.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. For *K* aligned TUs, the Manipuri-English transliteration model based on the Modified Joint Source Channel Model for transliteration (Ekbal et. al., 2006) is given by the following equations (1) and (2).

$$P(S,T) = \prod_{k=1}^{K} P(< s,t >_k \mid < s,t >_{k-1}) \, s_{k+1} \quad (1)$$

$$S \rightarrow T(S) = \arg\max_T \{P(T) \times P(S,T)\} \quad (2)$$

A TU correspondence <s, t> is called a transliteration pair of the source language S and the target language T. In this model, the previous TUs in both the source and target sides are considered as context. This model is extended to Manipuri since the Bengali script is used in Manipuri also.

5 SVM based RMWEs identification and Bilingual RMWE Extraction

The Support Vector Machine (SVM) based machine learning approach (Vapnik, 1995) works on discriminative approach and makes use of both positive and negative examples to learn the distinction between the two classes. The SVMs are known to robustly handle large feature sets and to develop models that maximize their generalizability. Consider a set of training data for a two-class problem: $\{(x_1, y_1),.....(x_N, y_N)\}$, where $x_i \varepsilon R^D$ is a feature vector of the i^{th} sample in the training data and $y_i \varepsilon \{+1, -1\}$ is the class to which x_i belongs. The goal is to find a decision function that accurately predicts class y for an input vector x.

A non-linear SVM classifier gives a decision function f (x) = sign ($g(x)$) for an input vector where,

$$g(x) = \sum_{i=1}^{m} w_i K(x, z_i) + b \qquad (3)$$

Here, f(x) = +1 means x is a member of a certain class and f(x) = -1 means x is not a member. The support vectors that are representatives of training examples are z_i and m is the number of support vectors. Therefore, the computational complexity of g(x) is proportional to m. Support vectors and other constants are determined by solving a certain quadratic programming problem. $K(x, z_i)$ is a *kernel* that implicitly maps vectors into a higher dimensional space. Typical kernels use dot products: $K(x, z_i) = k(x.z)$. A polynomial kernel of degree d is given by $K(x, z_i) = (1+x)^d$. We can use various kernels and the design of an appropriate kernel for a particular application is an important research issue.

The MNE/RMWE tagging system includes two main phases: training and classification. The training process has been carried out by YamCha[2] toolkit, an SVM based tool for detecting classes in documents and formulating the MNE/RMWE tagging task as a sequence labeling problem. Here, both one vs rest and pairwise multi-class decision methods have been used. Different experiments with the various degrees of the polynomial kernel function have been carried out. In one vs rest strategy, K binary SVM classifiers may be created where each classifier is trained to distinguish one class from the remaining K-1 classes. In pairwise classification, we constructed K (K-1)/2 classifiers considering all pairs of classes, and the final decision is taken by their weighted voting. For classification, the TinySVM-0.07[3] classifier has been used that seems to be one of the best optimized among publicly available SVM toolkits.

In the present work, the bilingual RMWEs are extracted automatically. The difficulty lies in how to integrate the bilingual MWEs into existing the SMT system to improve system performance. The RMWEs are identified from the Manipuri source side of the parallel corpora. The RMWEs are identified (Singh and Bandyopadhyay, 2010) using support vector machine (SVM) based machine learning technique. The various features used are context word, prefix, suffix, previous RMWE information, POS information, word length, digit and infrequent word features. The SVM based RMWE identification system shows recall, precision and F-Score of 94.62%, 93.53% and 94.07% respectively.

The target equivalents of the RMWEs are extracted by running the GIZA++ alignment tool followed be candidate translation extraction from the sentence pairs using the algorithm described in (Och, 2002). There are 25,921 RMWEs in the training corpus. Thus, an additional phrase table containing the automatically extracted RMWEs is constructed.

6 Bilingual NEs Extraction

Named Entities (NEs) are indentified using the SVM based Manipuri named entity recognition technique (Singh et al., 2010a) on news corpus. At present, 4649016 Manipuri wordforms has been collected from the website. The NE identification in Manipuri is difficult and challenging because (a) Manipuri is less privileged and resource constrained language, (b) unlike English, Manipuri lacks capitalization, (c) NEs can appear in the Manipuri dictionary as well and (d) Manipuri is highly agglutinative. The major NE tags are person name, location name, organization name and miscellaneous name. The identified Manipuri NEs are transliterated using the modified joint source channel techniques discussed in Section 4. There

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

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

are 529522 NEs identified and transliterated of which 284521 are multiword named entities (MNE). An additional phrase table of the bilingual NEs is constructed.

7 Transliterated non-NE list Preparation

Manipuri is influenced and enriched by the Indo-Aryan languages of Sanskrit origin and English. So the presence of transliterated English words in the Manipuri news corpus is significantly high and it is important to handle these words during the translation process at the appropriate step. So, a list of 2611 Manipuri words and their English transliterations is developed from the news corpus. The non-NE list consists of both single word as well as multiple words. An additional phrase table based on this non-NE transliteration list is built in order to integrate in the present system.

8 Integration of MWE

In the present task, we employed multiple phrase tables using the Moses decoder. One phrase table is trained from the parallel corpora and second one is built using the MWEs extracted using the techniques described in section 5 and 6. For simplicity, the probability 1 is assigned to all the four probabilities of the MWE phrase table. During the decoding process, the MWEs are searched in both the phrase tables.

One of the possible techniques of integrating MWE in the SMT system is by introducing a new feature in a phrase table that indicates the presence of MWE.

9 The SMT Systems

We developed the Manipuri-English SMT systems using the state-of-the-art Moses. The various features are combined in the log-linear model. The log linear model to obtain the best translation ê of the source sentence f is given by the equation below:

$$\hat{e} = \arg\max_{e} P(e|f) = \arg\max_{e} \sum_{i=1}^{n} \lambda_i h_i(e, f) \quad (4)$$

where, h_i and λ_i denote the i[th] feature function and weight respectively. The feature weight λ_i in the log linear model is determined by using the minimum error rate training method (Och, 2003). Intuitively, the $P(e|f)$ depend on *language model* — $P(e)$ and *translation model* — $P(f|e)$.

10 Experimental Setup

The first Manipuri-English SMT task is reported in (Singh and Bandyopadhyay, 2010c) on news domain using factored translation model demonstrating improvement not only in the BLEU and NIST scores but also improvement in the fluency and adequacy by subjective evaluation method. Earlier, an English-Manipuri SMT system using morpho-syntactic and semantic information is reported by (Singh and Bandyopadhyay, 2010d). In the present experimental setup, Moses decoder (Koehn et al., 2005) is used which can support multiple phrase tables. The target language model is developed using the SRILM (Stolcke, 2002) toolkit. The language model is the 4-gram model using Kneser-Ney smoothing (Kneser and Ney, 1995) of the target language news corpus collected from the www.thesangaiexpress.com. GIZA++ is used to build IBM Model 4. The minimum error rate training (Och, 2003) determine the feature weights on the development set.

Several experiments are conducted using the phrase tables built from the MWEs. There is also an experiment on the combination of all the phrase tables consisting of the baseline, RMWE, NEs and transliterated non-NEs.

11 Evaluation Results

The corpus statistics used in the experiment is given in the table below.

	Number of sentences	Number of words
Training	10350	296728
Development	600	16520
Test	500	15204

Table 4: Corpus Statistics

The automatic scoring metrics are useful for fast evaluation of higher number of test sentences. The

automatic evaluation scores are carried out using the BLEU (Papineni, 2002) with brevity penalty and NIST (Doddington, 2002). The BLEU and NIST automatic scores of various models are given in the table below:

Model	**BLEU**	**NIST**
Baseline	**13.452**	**4.31**
Baseline + RMWE	13.829	4.43
Baseline +NE	13.901	4.47
Baseline + Transliterated non-NE	13.911	4.21
Baseline + RMWE + NE + Transliterated non-NE	**15.023**	**5.21**

Table 5: Automatic Scores of Manipuri-English SMT systems

Level	**Interpretation**
4	Flawless with no grammatical error
3	Good output with minor errors
2	Disfluent ungrammatical with correct phrase
1	Incomprehensible

Table 6: Fluency scale

Level	**Interpretation**
4	Full meaning is conveyed
3	Most of the meaning is conveyed
2	Poor meaning is conveyed
1	No meaning is conveyed

Table 7: Adequacy scale

Subjective evaluation is carried out on 100 test sentences using the fluency and adequacy scales given in the Tables 6 and 7. The subjective evaluation is carried out by two bilingual human judges. The evaluation indicates the improvement in adequacy but not much on fluency as shown in Table 8. The case markers and the inflectional suffixes are not taken care of with special treatment despite the agglutinative behavior of Manipuri language towards RMWEs, NEs and non-NEs; hence the fluency is not addressed significantly.

Statistical significant test is performed to judge if a change in score that comes from a change in the system reflects a change in the overall translation quality. It is found that the difference between the baseline and the (Baseline + RMWE + NE + Transliterated non-NE) model is significant producing statistically significant improvements as measured by the bootstrap resampling method (Koehn, 2004) on BLEU.

	Manipuri Sentence length	**Fluency**	**Adequacy**
Baseline	<=15 words	1.93	2.31
	>15 words	1.51	1.76
Baseline+ RMWE	<=15 words	1.92	2.85
	>15 words	1.62	2.07
Baseline + NE	<=15 words	2.06	2.82
	>15 words	1.75	2.10
Baseline + Transliterated non-NE	<=15 words	2.10	2.79
	>15 words	1.67	2.10
Baseline + RMWE + NE + Transliterated non-NE	<=15 words	2.11	3.11
	>15 words	1.72	2.78

Table 8: Subjective Evaluation Scores

12 Conclusion and Future Work

The present work reports the tight integration of the RMWE, named entities (both single word and multiword) and transliterated non-NE entities into the Manipuri-English SMT task. The presence of these entities in the Manipuri news corpus is significantly high and their special treatment in the machine translation is absolutely necessary. In order to identify the language specific multiword expressions, the SVM based machine learning technique is utilized. Since the translation output of the SMT system is further propagated from the quality of the automatic RMWE and MNE identification using SVM based machine learning

technique and multiword extraction using GIZA++ alignment, it is very important that these steps are better addressed in an effective way with more training data in future. By integrating RMWEs, MNEs and non-NEs, the adequacy of the translation output is improved; however, there is no significant improvement in the fluency. The fluency can be better addressed by incorporating morphological information such as proper handling of the case markers of the MNEs. On the other hand, the BLEU and NIST scores are improved over the baseline from 13.452 and 4.31 to 15.023 and 5.21 respectively.

References

Carpuat, Marine, and Mona Diab. 2010. Task based Evaluation of Multiword Expressions: aPilot Study in Statistical Machine Translation. In Proceedings of Human Language Technology conference and the North American Chapter of the Association for Computational Linguistics conference (HLT-NAACL 2010), Pages 242-245, Los Angeles, CA.

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

Ekbal, Asif, Sudip Kumar Naskar, Bandyopadhyay, S. 2006. A Modified Joint Source-Channel Model for Transliteration, Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, Pages 191–198, Sydney.

Lambert, Patrik and Rafael Banchs. 2005. Data inferred multi-word expressions for statistical machine translation. In Proceedings of Machine Translation Summit X, Pages 396–403.

Koehn, Philipp and Knight, Kevin. 2003. Empirical Methods for Compound Splitting, Proceedings of the EACL 2003. Pages 187-194.

Koehn, Philipp. 2004. Statistical significance tests for machine translation evaluation. In EMNLP- 2004: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, 25-26 July 2004, Pages 388-395, Barcelona, Spain.

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, Pages 177-180, Prague, Czech Republic.

Och, Franz J., 2002, Statistical Machine Translation: From Single-Word Models to Alignment Templates. Ph.D. Thesis, Computer Science Department, RWTH, Aachen, Germany.

Och, Franz J., 2003. Minimum error rate training in Statistical Machine Translation, Proceedings of ACL.

Och, Franz J. and H. Ney. 2003. A systematic comparison of various statistical alignment models. Computational Linguistics, 29(1):19–51.

Pal, Santanu, Sudip Kumar Naskar , Pavel Pecina, Sivaji Bandyopadhyay, Andy Way, 2010. Handling Named Entities and Compound Verbs in Phrase-Based Statistical Machine Translation, In Proceedings of the Multiword Expressions: From Theory to Applications (MWE 2010), Pages 46–54, Beijing.

Papineni, K.A., Roukos, S., Ward, T., and Zhu, W.J., 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (ACL), Philadelphia,.

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, ACLIJCNLP 2009, Pages 47-54, Suntec, Singapore.

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 3th International Conference on Intelligent Text Processing and Computational Linguistics(CICLing-2002), Pages 1–15.

Singh, Thoudam Doren, Nongmeikapam Kishorjit, Asif Ekbal, Sivaji Bandyopadhyay, 2009. Named Entity Recognition for Manipuri using Support Vector Machine, In proceedings of 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC 23), Pages 811-818, Hong Kong.

Singh, Thoudam Doren and Sivaji Bandyopadhyay, 2010a. Web Based Manipuri Corpus for Multiword NER and Reduplicated MWEs Identification using SVM, Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing (WSSANLP), the 23rd International Conference on Computational Linguistics (COLING), Pages 35–42, Beijing.

Singh, Thoudam Doren and Sivaji Bandyopadhyay, 2010b. Semi Automatic Parallel Corpora Extraction from Comparable News Corpora, In the International Journal of POLIBITS, Issue 41 (January - June 2010), ISSN 1870-9044, Pages 11-17.

Singh, Thoudam Doren and Sivaji Bandyopadhyay, 2010c. Manipuri-English Bidirectional Statistical Machine Translation Systems using Morphology and Dependency Relations, *In Proceedings of Syntax and Structure in Statistical Translation* (SSST-4) *of 23rd International Conference on Computational Linguistics* (COLING), Pages 75-83, Beijing.

Singh, Thoudam Doren and Sivaji Bandyopadhyay, 2010d. Statistical Machine Translation of English-Manipuri using Morpho-Syntactic and Semantic Information, *In proceedings of Ninth Conference of the Association for Machine Translation in Americas* (AMTA 2010), Pages 333-340, Denver, Colorado, USA.

Singh, Ch. Yashawanta, 2000, Manipuri Grammar, *Rajesh Publications*, Pages 190-204, Delhi.

Vapnik, Vladimir N. 1995: The nature of Statistical learning theory. Springer