Web Based Manipuri Corpus for Multiword NER and Reduplicated MWEs Identification using SVM

Thoudam Doren Singh
Department of Computer Science and Engineering
Jadavpur University
thoudam.doren@gmail.com

Sivaji Bandyopadhyay
Department of Computer Science and Engineering
Jadavpur University
sivaji_cse_ju@yahoo.com

Abstract

A web based Manipuri corpus is developed for identification of reduplicated multiword expression (MWE) and multiword named entity recognition (NER). Manipuri is one of the rarely investigated language and its resources for natural language processing are not available in the required measure. The web content of Manipuri is also very poor. News corpus from a popular Manipuri news website is collected. Approximately four and a half million Manipuri wordforms have been collected from the web. The mode of corpus collection and the identification of reduplicated MWEs and multiword NE based on support vector machine (SVM) learning technique are reported. The SVM based reduplicated MWE system is evaluated with recall, precision and F-Score values of 94.62%, 93.53% and 94.07% respectively outperforming the rule based approach. The recall, precision and F-Score for multiword NE are evaluated as 94.82%, 93.12% and 93.96% respectively.

1 Introduction

The NER and MWE identification are important tasks for natural language applications that include machine translation and information retrieval. The present work reports the NER and reduplicated MWE identification of Manipuri on web based news corpus. The use of web as a corpus for teaching and research on languages has been proposed several times (Rundell, 2000; Fletcher, 2001; Robb, 2003; Fletcher 2004). A special issue of the Computational Linguistics journal on web as corpus (Kilgarriff and Grefenstette, 2003) was published. Several studies have used different methods to mine web data. The web walked into the ACL meetings starting in 1999. The special interest group of ACL on web as corpus is promoting interest in the use of the web as a source of linguistic data, and as an object of study in its own right. India is a multilingual country with a lot of cultural diversity. Bharati et al. (2001) reports an effort to create lexical resources such as transfer lexicon and grammar from English to several Indian languages and dependency Treebank of annotated corpora for several Indian languages. In Indian context, a web based Bengali corpus development work from web is reported in Ekbal and Bandyopadhyay (2008) and Manipuri-English semi automatic parallel corpora extraction by Singh et. al., (2010). Newspaper is a huge source of readily available documents. In the present work, the Manipuri monolingual corpus has been developed from web for NLP and related tasks.

Manipuri is a scheduled Indian language spoken approximately by three million people mainly in the state of Manipur in India and in the neighboring countries namely Bangladesh and Myanmar. It is a Tibeto-Burman language and highly agglutinative in nature, 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. In Manipuri, words are formed in three processes called affixation, derivation and compounding. The majority of the roots found in the

Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing (WSSANLP), pages 35–42,
the 23rd International Conference on Computational Linguistics (COLING), Beijing, August 2010

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 etc. are not yet available in Manipuri in the required measure.

In the present work, the tasks of identification of Manipuri multiword named entity (MNE) and reduplicated multiword expression (RMWE) identification using support vector machine (SVM) learning technique on the corpus collected from web is reported.

Works on multiword expressions (MWEs) have started with a momentum in different languages. In the Indian context, some of the works can be seen in (Dandapat et. al., 2006; Kunchukuttan and Damani, 2008; Kishorjit et. al., 2010). The identification of MWEs in several languages concentrate on compound nouns, noun-verb combination, some on idioms and phrases and so on 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 are quite large in number in Manipuri and there are wide varieties of reduplicated MWEs in Manipuri.

2 Manipuri News Corpus and Statistics

The content of Manipuri language on the web is very poor. One of the sources is the daily news publications. Again, there is no repository. Thus, the possibility of deploying web crawler and mining the web corpus is not possible. The Manipuri news corpus is collected from http://www.thesangaiexpress.com/ covering the period from May 2008 to May 2010 on daily basis. The Manipuri news is available in PDF format. A tool has been developed to convert contents from PDF documents to Unicode format. There are 15-20 articles in each day. Considering the Manipuri corpus covering the period from May 2008 to May 2010, there are 4649016 wordforms collected[1].

2.1 Conversion from PDF to UTF-8

The general Manipuri news collected is in PDF format. A tool has been developed to convert Manipuri news PDF articles to Bengali Unicode. 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 738 entries. There are conjuncts of 2, 3 and 4 characters. Sub-tables for each of the conjuncts are prepared. English words are present on the Manipuri side of the news and they are filtered to avoid unknown character features.

2.2 Use of language resources

The Manipuri web corpus collected from the web is cleaned by removing the unknown characters. After the cleaning process, the running texts are picked up followed by spelling correction. The web based news corpus is POS tagged using the 26 tagset[2] defined for the Indian languages based on the work of (Singh et. al., 2008). The Manipuri news corpus developed in this work has been used to identify MNE and RMWEs identification.

3 Support Vector Machine

The SVM (Vapnik, 1995) is based 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. Suppose we have a set of training data for a two-class problem: $\{(x_1, y_1), ...(x_N, y_N)\}$, where $x_i \varepsilon$ RD 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

[1] There are no publications on some occasions.

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

for an input vector x. A non-linear SVM classifier gives a decision function $f(x)$= sign ($g(x)$) for an input vector where, $g(x) = \sum_{i=1}^{m} w_i K(x, z_i) + b$ 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 vector is represented by z_i and stands for the training examples; 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[3] 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 given by their weighted voting. For classification, the TinySVM-0.07[4] classifier has been used that seems to be the best optimized among publicly available SVM toolkits.

4 Multiword Named Entity Recognition

Named Entity Recognition for Manipuri is reported in (Singh et. al., 2009). The present work focuses and reports on the recognition of multiword NEs. In order to identify the MNEs, 28,629 wordforms from Manipuri news corpus has been manually annotated and used 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 MNE classes.

MNE 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 MNEs.
- A lot of MNEs 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 MNEs can appear in subject and object positions making the NER task more difficult compared to others.
- Manipuri is a resource-constrained language. Annotated corpus, name dictionaries, sophisticated morphological analyzers, POS taggers etc. are not yet available.

MNE Tag	**Meaning**	**MNE Examples**
B-LOC	Beginning, Internal or the End of a multiword location name	থাঙ্গা (Thanga)
I-LOC		মোইরাংথেম (Moirangthem)
E-LOC		লৈকায় (Leikai)
B-PER	Beginning, Internal or the End of a multiword person name	ওইনাম (Oinam)
I-PER		ইবোবি (Ibobi)
E-PER		মীতৈ (Meetei)

Table 1. Named entity examples

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

In the present work, the NE tagset used have been further subdivided into the detailed categories in order to denote the boundaries of MNEs properly. Table 1 shows examples.

5 Reduplicated MWEs Identification

Manipuri is very rich in RMWEs like other Tibeto-Burman languages. The work of (Singh, 2000) describes the linguistic rules for identifying reduplicated words. A rule based Manipuri RMWE identification is reported in (Kishorjit and Bandyopadhyay, 2010). 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: 1) Complete RMWEs, 2) Partial RMWEs, 3) Echo RMWEs and 4) 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 clause is repeated once forming a single unit regardless of phonological or morphological variations.

মরিক মরিক ('*marik marik*') which means 'drop by drop'.

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

করি করি ('*kari kari*') means 'what/which'.

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, চথোক চৎসিন ('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 reduplication 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 and so also is the MWE. 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'.

5.1 Role of suffix and prefix

Apart from the above cases meaningful prefixes or suffixes are used with RMWEs otherwise they are ungrammatical.

Suffixes/ wh- duplicating words	Part of Speech
দা (*–da*), গি (*–gi*) and কি (*–ki*)	Noun
বা (*–ba*) and পা (*–pa*)	Adjective
না (*–na*)	Adverb
করি করি ('*kari kari*'), কনা কনা ('*kanaa kanaa*'), কদায় কদায় ('*kadaay kadaay*') and করম করম ('*karam karam*')	Wh- question type

Table 2. Suffixes/wh- duplicating words list used in Complete Reduplication and POS tagging

Prefix: With such prefixes the semantic shapes are different and sometimes even the same prefix carries a different meaning. By these prefixation, the root is reduplicated as given below:

{[ই(*i*)-/পঙ(*pang*)-/খঙ(*khang*)-/ত(*ta*)-/পুম(*pum*)-/শুক(*suk*)] + Root }

→

{[ই(*i*)-/পঙ(*pang*)-/খঙ(*khang*)-/ত(*ta*)-/পুম(*pum*)-/শুক(*suk*)] + Root + Root}

মহাক্না ইৱাঙ ৱাঙই
mahaak-na i-waang waang-ngi
he/she-nom –tall tall-asp
He/She is the tallest

Suffix: There are some suffixes that carry certain meaning when used with RMWEs. Commonly used suffixes are, ত্রিক (-trik) / দ্রিক (-drik), থ্রোক (-throk), দ্রোং (-drong), শুক (-suk), শঙ (-sang), দ্রিং (-dring), শিত্ (-sit), শিন (-sin), দ্রেং (-

dreng), শোক (-sroke) etc. Generally these suffixes indicate a superlative degree or emphatic meaning.

Some examples are as follows,

মুনত্রিক মুনবা

mun-trik mun-ba

ripe ripe

'very ripe'

Role of affix in Partial Reduplication: Character-wise comparisons are done with not less than two characters either from front or rear for both the words since the second word is not a complete repetition.

Also the last few characters of the first word and the same number of first characters of the second word are compared to check the partial reduplication. The prefixes or suffixes are verified with a list of accepted suffixes and prefixes (see table 2) to validated the reduplication.

Role of affix in Echo Reduplication: Identification of echo reduplication is done by comparing the equality of suffixes of consecutive two words w1 and w2.

6 Best Feature Selection for SVM

The use of prefix/suffix information works well for the highly inflected languages like the Indian languages. Different combinations from the following set for identifying the best feature set for MNE/RMWE are experimented: F={ $w_{i-m}, .., w_{i-1}, w_i, w_{i+1}, ..., w_{i+n}$, |prefix|<=n, |suffix|<=n, MNE/RMWE tag(s) of previous word(s), POS tag(s) of the current and/or the surrounding word(s), First word, Length of the word, Digit information, Infrequent word}, where w_i is the current word; w_{i-m} is the previous m^{th} word and w_{i+n} is the next n^{th} word. Following are the details of the features:

1 Context word feature: Preceding and following words of a particular word since the surrounding words carry effective information for the identification of MNE/RMWEs.

2 Word suffix: Word suffix information is helpful to identify MNE/RMWEs. This is based on the observation that the MNE/RMWEs share some common suffixes. The fixed length (say, n) word suffix of the current and/or the surrounding word(s) can be treated as the feature. If the length of the corresponding word is less than or equal to n − 1 then the feature values are not defined and are denoted by 'ND'. The feature value is also not defined (ND) if the token itself is a punctuation symbol or contains any special symbol or digit. Word suffixes are the effective features and work well for the highly inflective Indian languages like Manipuri.

3 Word prefix: Word prefixes are also helpful to identify MNE/RMWEs. It is based on the observation that MNE/RMWEs share some common prefix strings. This feature has been defined in a similar way as that of the fixed length suffixes.

4 MNE and RMWE Information: The MNE/RMWE tag(s) of the previous word(s) have been used as the only dynamic feature in the experiment. The output tag of the previous word is very informative in deciding the MNE/RMWE tag of the current word.

5 Digit features: Several binary valued digit features have been defined depending upon the

(i). Presence and/or the exact number of digits in a token.

(a). CntDgtCma: Token consists of digits and comma

(b). CntDgtPrd: Token consists of digits and periods

(ii). Combination of digits and symbols. For example,

(a). CntDgtSlsh: Token consists of digit and slash

(b). CntDgtHph: Token consists of digits and hyphen

(c). CntDgtPrctg: Token consists of digits and percentages

(iii). Combination of digit and special symbols. For example,

(a). CntDgtSpl: Token consists of digit and special symbol such as $, # etc.

These binary valued digit features are helpful in recognizing miscellaneous NEs such as measurement expression and percentage expression.

6 Infrequent word: The frequencies of the words in the training corpus have been calculated. A cut off frequency has been chosen in order to consider the words that occur with less than the cut off frequency in the training corpus. A binary valued feature 'Infrequent' is defined to check whether the current word appears in this infrequent word list or not. This is based on the observation that the infrequent words are most probably MNE/RMWEs.

7 Length of a word: This binary valued feature is used to check whether the length of the current word is less than three or not. We have observed that very short words are most probably not the MNE/RMWEs.

8 Part of Speech (POS) information: We have used an SVM-based POS tagger (Singh et. al., 2008) that was originally developed with 26 POS tags, defined for the Indian languages. The POS information of the current and/or the surrounding words can be effective for MNE/RMWE identification.

The Table 3 gives the statistics of training, development and test sets. The various notations used in the experiments are presented in Table 4. The Table 5 shows the recall (R), precision (P) and F-Score (FS) in percentage in the development set.

	Training	**Development**	**Test**
# of sentences	1235	732	189
#of wordforms	28,629	15,000	4,763
# of distinct wordforms	8671	4,212	2,207

Table 3. Statistics of the training, development and test sets

Notation	**Meaning**
W[-i,+j]	Words spanning from the i^{th} left position to the j^{th} right position
POS[-i, +j]	POS tags of the words spanning from the i^{th} left to the j^{th} right positions
Pre	Prefix of the word
Suf	Suffix of the word
NE [-i, -j]	NE tags of the words spanning from the i^{th} left to the j^{th} left positions

Table 4. Meaning of the notations

Feature	**R %**	**P %**	**FS %**
Static: W[-2,+2], POS[-2,+2], |Pre|<=3, |Suf|<=3, Length, Infrequent, FirstWord, Digit **Dynamic:** MNE/RMWE[-2,-1]	**94.26**	**96.72**	**95.47**
Static: W[-3,+3], POS[-3,+3], |Pre|<=3, |Suf|<=3, Length, Infrequent, FirstWord, Digit **Dynamic:** MNE/RMWE[-3,-1]	88.23	94.82	91.40
Static: W[-3,+2], POS[-3,+2], |Pre|<=3, |Suf|<=3, Length, Infrequent, FirstWord, Digit **Dynamic:** MNE/RMWE[-3,-1]	90.32	93.18	91.72
Static: W[-4,+3], POS[-4,+3], |Pre|<=3, |Suf|<=3, Length, Infrequent, FirstWord, Digit **Dynamic:** MNE/RMWE[-2,-1]	88.15	92.62	90.32
Static: W[-4,+3], POS[-4,+3], |Pre|<=3, |Suf|<=3, Length, Infrequent, FirstWord, Digit **Dynamic:** MNE/RMWE[-3,-1]	86.24	92.31	89.17
Static: W[-2,+2], POS[-	88.	91.	90.

2,+2], |Pre|<=4, |Suf|<=4, Length, Infrequent, FirstWord, Digit **Dynamic:** MNE/RMWE[-2,-1]	70	49	07
Static: W[-3,+3], POS[-3,+3], |Pre|<=4, |Suf|<=4, Length, Infrequent, FirstWord, Digit **Dynamic:** MNE/RMWE[-3,-1]	85.05	90.09	87.49
Static: W[-4,+3], POS[-4,+2], |Pre|<=4, |Suf|<=4, Length, Infrequent, FirstWord, Digit **Dynamic:** MNE/RMWE[-2,-1]	78.55	89.54	83.68
Static: W[-4,+4], POS[-4,+4], |Pre|<=4, |Suf|<=4, Length, Infrequent, FirstWord, Digit **Dynamic:** MNE/RMWE[-3,-1]	73.71	89.44	80.81

Table 5. Results on the development set

7 Results on the Test Set

The best feature set (F) of Manipuri MNER and RMWE 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]. After the selection of the best feature set, the SVM based system for MNE and RMWEs is tested on the test set of 4,763 wordforms.

Reduplicated MWE type	**Recall %**	**Precision %**	**F-Score %**
Complete and mimic	96.21	95.12	95.66
Partial	88.32	85.03	86.64
Echo	97.76	96.45	97.10
Double	93.23	94.23	93.72
Semantic	74.45	81.56	77.84

Table 6. Result on RMWE test set

In this work, SVM that parses from left to right is considered. The break-up of the RMWEs and the scores are given in Table 6. The handling of semantic RMWEs requires further investigation to improve the performance. The rule based RMWE identification (Kishorjit and Bandyopadhyay, 2010) shows a recall, precision and F-Score of 94.24%, 82.27% and 87.68% respectively.

Multiword NE	**Recall %**	**Precision %**	**F-Score %**
Person	94.21	95.12	94.66
Location	94.32	95.03	94.67
Organization	95.76	93.45	94.59
Miscellaneous	92.23	91.23	91.72

Table 7. Result on MNE test set

It is observed that the SVM based system outperforms the rule based system. Table 7 shows the break-up scores of different types of MNEs and Table 8 shows the overall scores of MNE and RMWE.

	Recall %	**Precision %**	**F-Score %**
MNE	94.82	93.12	93.96
RMWE	94.62	93.53	94.07

Table 8. Overall recall, precision and F-Scores on test set

8 Conclusion

In this paper, the development of RMWEs identification and recognition of MNE for a resource-constrained language using web based corpus of Manipuri is reported. This training data of 28,629 is then manually annotated with a coarse-grained tagset of four NE tags and six RMWEs in order to apply SVM and tested with 4,763 wordforms. The SVM classifier makes use of the different contextual information of the words along with the various orthographic word-level features. A number of experiments have been carried out to find out the best set of features for MWE in Manipuri. The SVM based RMWE system outperforms the rule based approach. The SVM based RMWE shows recall, precision and F-Score of 94.62%, 93.53% and 94.07% respectively. The rule based RMWE

identification shows a recall, precision and F-Score of 94.24%, 82.27% and 87.68% respectively. The overall performance of the system shows reasonable output for both MNE and RMWE.

References

Bharati, A., Sharma, D. M., Chaitanya, V., Kulkarni, A. P., & Sangal, R., 2001. LERIL: Collaborative effort for creating lexical resources. *In Proceedings of the 6th NLP Pacific Rim Symposium Post-Conference Workshop*, Japan.

Dandapat, S., Mitra, P., and Sarkar, S., 2006. Statistical investigation of Bengali noun-verb (N-V) collocations as multi-word-expressions, *In Proceedings of Modeling and Shallow Parsing of Indian Languages (MSPIL)*, Mumbai, pp 230-233

Ekbal, A., and Bandyopadhyay, S., 2008. A web based Bengali news corpus for Named Entity Recognition, *Lang Resources & Evaluation* (2008) 42:173–182, Springer

Fletcher, W. H., 2001. Concordancing the web with KWiCFinder. *In Proceedings of the Third North American Symposium on Corpus Linguistics and Language Teaching*, Boston, MA, 23–25 March 2001.

Fletcher, W. H., 2004. Making the web more use-ful as source for linguists corpora. In U. Conor & T. A. Upton (Eds.), *Applied corpus linguists: A multidimensional perspective* (pp. 191–205). Amsterdam: Rodopi.

Kilgarriff, A., and Grefenstette, G., 2003. Introduction to the special issue on the web as corpus. *Computational Linguistics*, 29(3), 333–347.

Kishorjit, N., and Bandyopadhyay, S., 2010. Identification of Reduplicated MWEs in Manipuri: A Rule Based Approch, *In proceedings of 23rd International Conference on the Computer Processing of Oriental Languages (ICCPOL 2010) - New Generation in Asian Information Processing* , Redmond City, CA

Kunchukuttan, A., and Damani, O. P., 2008. A System for Compound Nouns Multiword Expression Extraction for Hindi, *In Proceedings of 6th International conference on Natural Language Processing (ICON 2008)*, Pune, India

Robb, T., 2003. Google as a corpus tool? *ETJ Journal*, 4(1), Spring.

Rundell, M., 2000. The biggest corpus of all. Humanising Language Teaching, 2(3)

Singh. Chungkham Y., 2000. Manipuri Grammar, *Rajesh Publications*, Delhi, pp 190-204

Singh, Thoudam D., Ekbal, A., Bandyopadhyay, S. 2008. Manipuri POS tagging using CRF and SVM: A language independent approach, *In proceeding of 6th International conference on Natural Language Processing (ICON -2008)*, Pune, India, pp 240-245

Singh, Thoudam D., Kishorjit, N., Ekbal, A., Bandyopadhyay, S., 2009. Named Entity Recognition for Manipuri using Support Vector Machine, *In proceedings of 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC 23)*, Hong Kong, pp 811-818

Singh, Thoudam D., Singh, Yengkhom R. and Bandyopadhyay, S., 2010. Manipuri-English Semi Automatic Parallel Corpora Extraction from Web, *In proceedings of 23rd International Conference on the Computer Processing of Oriental Languages (ICCPOL 2010) - New Generation in Asian Information Processing* , Redmond City, CA

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