Addressing some Issues of Data Sparsity towards Improving English-Manipuri SMT using Morphological Information

Thoudam Doren Singh
Centre for Development of Advanced Computing (CDAC)
Gulmohor, Cross Road 9
Juhu, Mumbai-400049, India
thoudam.doren@gmail.com

Abstract

The performance of an SMT system heavily depends on the availability of large parallel corpora. Unavailability of these resources in the required amount for many language pair is a challenging issue. The required size of the resource involving morphologically rich and highly agglutinative language is essentially much more for the SMT systems. This paper investigates on some of the issues on enriching the resource for this kind of languages. Handling of inflectional and derivational morphemes of the morphologically rich target language plays important role in the enrichment process. Mapping from the source to the target side is carried out for the English-Manipuri SMT task using factored model. The SMT system developed shows improvement in the performance both in terms of the automatic scoring and subjective evaluation over the baseline system.

1 Introduction

Since the dawn of SMT system in the early 90s with the seminal work at IBM (Brown et al., 1992; Brown et al., 1993), there has been growth in the number of the SMT system for several language pairs. Performant SMT systems for the major languages are available. In the same time, such development is limited for less privileged and resource poor languages. Developing English-Manipuri SMT systems is one of such examples. Manipuri is a morphologically rich and highly agglutinative in nature. New words are easily coined by combination of various morphemes. Verb morphology is more complex and productive than noun morphology. In Manipuri, adjective and adverbs come from verbal root through derivational morphology. Aspectual marker goes with the derived forms. This language contains abundant reduplicated multiword expressions (RMWE). Language resource for this language pair is not available in the required measure.

2 Related Work

Several SMT systems between English and morphologically rich languages are reported. (Oflazer and El-Kahlout, 2007) investigated different representational granularities for sub-lexical representation in statistical machine translation work from English to Turkish by exploring different representational units in English to Turkish SMT. (Yeniterzi and Oflazer, 2010) further reported syntax-to-morphology mapping in factored phrase-based Statistical Machine Translation (Koehn and Hoang, 2007) from English to Turkish relying on syntactic analysis on the source side (English) and then encodes a wide variety of local and non-local syntactic structures as complex structural tags which appear as additional factors in the training data. On the target side (Turkish), they only perform morphological analysis and disambiguation but treat the complete complex morphological tag as a factor, instead of separating morphemes. Some of the SMT systems between English and morphologically rich languages which used morphemes to address the data sparsity are discussed below.

English-to-Czech phrase-based machine translation experiment (Bojar, 2007) with additional annotation of input and output tokens (multiple factors) used to explicitly model morphology by setting up various multiple factors and the amount of information in the morphological tags resulted in significant translation quality increase. Further, two contributions using factored phrase based model and a probabilistic tree transfer mode at deep syntactic layer are made by (Bojar and Hajič, 2008) of English-to-Czech SMT system. (Toutonova et al., 2007) reported the improvement of an SMT by applying word form prediction models from a stem using extensive morphological and syntactic information from source and target languages. (Gandhe et al., 2011) proposed a solution to augment the phrase table with all possible forms of a verb for improving the overall accuracy of the English–Hindi MT system by using simple stemmer and easily available monolingual data to generate new phrase table entries that cover the different variations seen for a verb. (Habash, 2008) presented four techniques for online handling of Out-of-Vocabulary words in Arabic-English Phrase based Statistical Machine Translation by using spelling expansion, morphological expansion, dictionary term expansion and proper name transliteration to reuse or extend a phrase table.

3 Enriching the Language Resource

SMT systems demand a large parallel corpus as training data. One of the approaches to develop SMT systems for less privileged and resource poor language is to enrich the resource through morphological processing by learning the general rules of morphology. This helps to increase the training data size and increase the coverage of the occurrence of the different words. The representation of words, i.e. the spelling has the most important role to play. Including different forms of a word in the training data makes a sense to improve the translation quality. While exploring the different representational units from English-to-Manipuri, there is lack of information at the source side for derivation and inflection of the target words. The verb morphology is more complex and productive. Focusing on the verb, the derivational morphology is more productive than the inflectional. Separating lemma, inflections and derivational morphemes allows the system to learn more about the different possible word formations. Thus, considering all the possible morpheme combination helps to enrich the language in order to cover more vocabulary.

3.1 Examples of Word Level Alignment between English and Manipuri

Consider the following example of English to Manipuri translation depicting the word level alignment by figure 1, stems and separated affixes by Figure 2 and phrasal level alignment by figure 3. From the example we can see that for each Manipuri word, there is a corresponding chunk of English (i.e., a group of words). This prompts us to correlate a group of English words with the help of a chunker and a Manipuri word.

But the question is to sort out how many possible morphemes are there to represent a word. Consider the following Manipuri sentence showing the possible translations due to non-standardized spelling.

English:

They came.

Manipuri Translation: (different possible orthographic and phonographic variations)

মখোয় লাক্লম্মী ।	(*makhoy lak-lammee*)
মখোয় লাকঅম্মী ।	(*makhoy lak-ammee*)
মখোয় লাক্লমঈ ।	(*makhoy lak-lam-ee*)
মখোয় লাক্লমই ।	(*makhoy lak-lam-i*)
মখোই লাক্লমই ।	(*makhoi lak-lam-i*)

Enrichment of the language resource is carried out for this languages category using morphemes with various orthographic variations. Inclusion of these various orthographic variations is absolutely necessary. However, the problem crops up how much of the training data is really increased in terms of size and improved in terms of alignment quality? When they are not handled in one of these steps these words become out-of-vocabulary (OOV) words. So, spelling expansion of these words in order to extend the phrase table becomes essential. This helps in the orthographic normalization and minimizing the data sparsity. The spelling expansion and morphological expansion are definitely helping minimizing data sparsity.

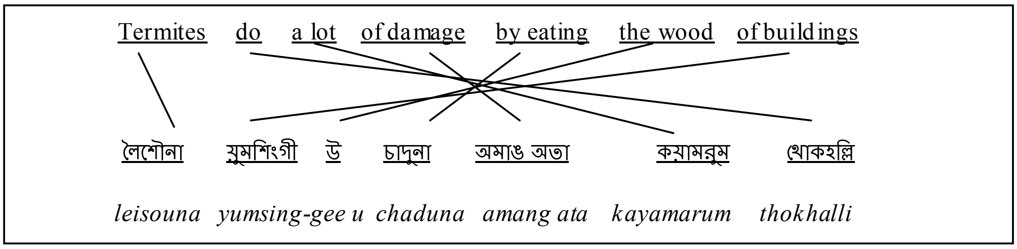

Figure 1: Word and Chunk level alignment between English and Manipuri

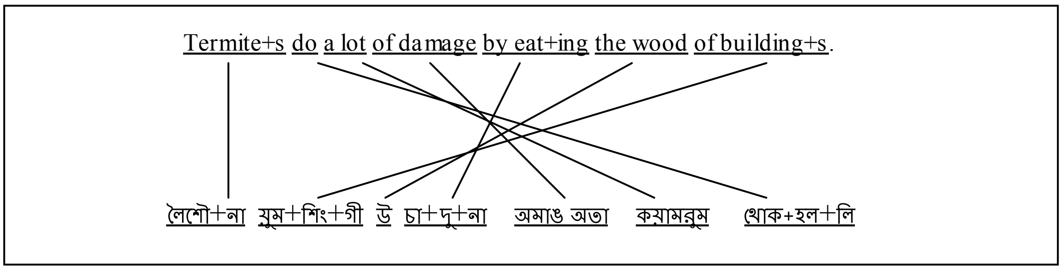

Figure 2: Stems and Morphemes separated between English and Manipuri alignment

(NP (NNS Termites)) → লৈশৌ+না

(VP (VBP do) → থোকহল+লি

(NP (DT a) (NN lot)) → কয়ামরুম

(PP (IN of)

(NP (NN damage))) → অমাঙ অতা

(PP (IN by)

(VP (VBG eating) → চা+দু+না

(NP (DT the) (NN wood)) → উ

(PP (IN of)

(NP (NNS buildings))) → য়ুম+শিং+গী

Figure 3: Phrasal Level Alignment between English and Manipuri

4 Key Aspects of Manipuri Morphology

In this agglutinative language the numbers of verbal suffixes are more than that of the nominal suffixes (Singh, 2000). New words are easily formed in Manipuri using morphological rules. 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). Manipuri morphological processing works are reported by (Singh and Bandyopadhyay, 2006) and (Singh and Bandyopadhyay, 2008).

4.1 Verb morphology

Three derivational categories may optionally precede the final inflectional suffix. The 1st LD suffixes signal adverbial meanings, the 2nd LD suffixes indicate evidentiality, the deitic reference of a verb, or the number of persons performing the action and the 3rd LD suffixes signal aspect and mood. Verb roots may also be used to form verbal nouns, adjectives and adverbs. Verbal nouns are formed through the suffixation of the nominalizer পা –pə to the verb root. The following is the list of word structure rules for verbs (Shobhana, 1997)

a. Verb → STEM INFL
b. STEM → Stem (3rd LD)
c. Stem → Stem (2nd LD)
d. Stem → Root (1st LD)
e. ROOT → root (root)
f. 3rd LD → (mood1)(mood2)(aspect)
g. 2nd LD → (2nd LD1),(2nd LD2),(2nd LD3)..
h. 1st LD → 1st LD

Derivational Prefixation	Root	1st LD	2nd LD	3rd LD	Inflection

Figure 4: General form of Verb Morphology

There are 3 categories (mood1, mood2, and aspect) belonging to the third level derivational (3rd LD) markers. The general form of verb morphology is shown in figure 4.

The sub-categorization frames of affixes will restrict that only nominal affixes occur with a noun and verbal affixes occur with a verb root. The derivational suffix order of the word চেকখাইরক্কনি (It'll get cracked) is given below:-

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

The রক *-rək* has allomorph লক-*lək.* রক *-rək* occurs after vowels while লক-*lək* occurs after consonants,

চারকএ *–ca-rək-ey* (ate there and came here)
চাম্লকএ *– cam-lək-ey* (washed there and came here)

The formation of verb can be of the form

Verb stem + aspect/mood → verb
থক *-thək* (drink) + লে *-le-* → থকলে *-thəkle* (has drunk)

The verbal noun is formed with the rule as given as

Verb Stem + Nominalizer → Verbal noun
থোং *-thong* (cook)+ বা *-ba* → থোংবা *-thongba* (to cook)

4.2 Noun Morphology

The following is the list of word structure rules for nouns (Shobhana, 1997)

N → Stem INFL (ENC)
Stem → stem (2nd LD)
Stem → ROOT(1st LD)
ROOT → (prefix) root (root)

Figure 5 shows the general form of noun morphology in Manipuri. Examples of some singular/plural noun forms are listed in Figure 6.

Pronominal prefix	Root	gender	number	Quantifier	Case

Figure 5: General form of Noun Morphology

Singular Form	Plural Form
উচেক -Uchek (bird)	উচেকশিং -Ucheksing(birds)
ম -Ma (He/She)	মখোয় -Makhoy (they)
মী -Mi (man)	মীয়াম -Mi-yaam (men)

Figure 6: Singular/Plural forms

Although case markers are functionally inflectional, they exhibit the clitic like characteristic of docking at the edge of a phrase. The word structure of rules of verbs and nouns are identical except for the category of the word level node, the possible terminal elements of the derivational and inflectional categories and the lack of the third level nominal derivation. Two examples to demonstrate the noun morphology are given below:-

মচানুপীশিংনা (*mə-ca-nu-pi-siŋ-nə*) ’by his/her daughters’
মচানুপাশিংনা (*mə-ca-nu-pa-siŋ- nə*) ’by his/her sons’

The ম *-mə* ‘his/her’ is the pronominal suffix and চা *-ca* ‘child’ is the noun root. The নু *-nu* ‘human’ is suffixed by পী *-pi* to indicate a female human and পা–*pa* to indicate a male human. শিং *–siŋ* or খোই *-khoy* or য়াম–yaam can be used to indicate plurality. *-siŋ* cannot be used with pronouns or proper nouns and *-khoy* cannot be used with nonhuman nouns. না *-nə* meaning ‘by the’ is the instrumental case marker.

4.3 Adjectives and Adverbs

In Manipuri, adjective and adverbs come from verbal root (in the example: ফ (ph)) through derivational morphology. Aspectual marker goes with the derived forms. Some of the examples are given as:

a)
The player is good
শান্নরোয়দু ফৈ
shannaroydu phei

b)
The player is still good
শান্নরোয়দু হৌজিকসু ফরি
shannaroydu houjiksu phari

c)
The player is always good
শান্নরোয়দু অদুম ফৈ
shannaroydu adum phei

d)
The player was good
শান্নরোয়দু ফরম্মী
shannaroydu pharami

5 Experiments

Indian languages are morphologically rich and have relatively free-word order where the grammatical role of content words is largely determined by their case markers and not just by their positions in the sentence. SMT systems between English and morphologically rich and highly agglutinative languages suffer badly if adequate training and language resource is not available and the accountability of individual morpheme is not considered. Machine Translation systems of Manipuri (the first Tibeto-Burman language for which MT system is developed) and English are reported by (Singh and Bandyopadhyay, 2010b) on developing the first Manipuri to English example based machine translation system followed by (Singh and Bandyopadhyay, 2010c) on development of English-Manipuri SMT system using morpho-syntactic and semantic information where the target case markers are generated based on the suffixes and semantic relations of the source sentence. Further (Singh and Bandyopadhyay, 2011a) reported on the development of bidirectional SMT system for English-Manipuri language pair using dependency relations, morphological information and parts of speech tags and (Singh and Bandyopadhyay, 2011b) continued reporting on the integration of reduplicated multiword expression and named entities into the English-Manipuri SMT system. In the present work, detailed morphological information such as the inflection and derivational morphemes are integrated into the system. In an effort to optimize, the training data, we experimented on the variation in the performance while making choice of sentence length for training. Being a highly agglutinative language, the translation performance is largely affected for longer training. The English-Manipuri parallel corpus on news domain developed by (Singh and Bandyopadhyay, 2010a) is used in the experiment. Two different models are developed as shown in figure 7; (i) inflectional model and (ii) inflectional + derivational model. The inflectional model uses lemma and suffix factors on the source side, lemma and suffix on the target side for lemma to lemma and suffix to suffix translations with generation step of lemma plus suffix to surface form. For the second model, two important treatments of the noun (Thoudam, 1982) and verb morphology for mapping with the corresponding source side is carried out considering the stem, derivational morphology and inflectional morphology of the target side. The reason why Manipuri inflectional morphology is to be treated as separate factor is that – it is comparatively easier to map to English dependency relations and suffix information (Singh and Bandyopadhyay, 2010c) to address the crux of the fluency. The BLEU score of this SMT system on the same corpus statistics as given in table 1 is 16.873 as reported earlier. Again, considering all Manipuri derivational morphology as one factor mapped to the complete phrase of the corresponding English phrases helps to cover overall meaning. This exercise reduces the overall burden to deal with individual morphemes with its discontinuous representation of English counterpart. We augment the training phrase table with 5000 manually prepared variants of verbs and nouns phrases for improving the overall accuracy of the SMT system. Manipuri uses Bengali script to represent the text. The wide variations of tone are not captured during the textual representation. Lexical ambiguity is very common in this language. This has resulted towards the requirement of a word sense disambiguation module. As part of this ongoing experiment, an additional lexicon of 11000 entries between English and Manipuri is employed to handle bits of sense disambiguation with the help of a word-based language model. The English sentences are processed with morpha (Minnen, 2003) and Stanford Parser (de Marneffe and Manning, 2008) is used for parsing. All words in Manipuri are bound except noun. We process the Manipuri corpus by segmenting into three parts, viz, (a) lemma (b) derivational morphemes and (c) inflections. The Manipuri stemmer (Singh and Bandyopadhyay, 2008) is used to separate the stem, derivational morpheme and inflections.

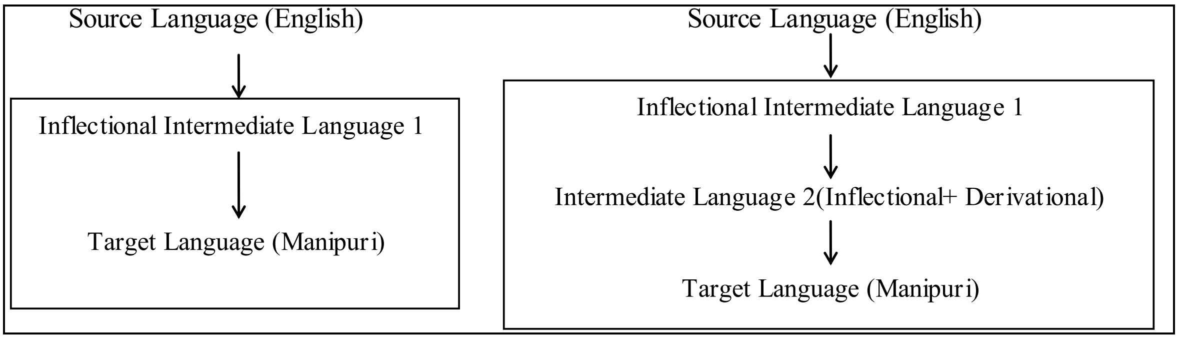
Source Language (English)
Inflectional Intermediate Language 1
Target Language (Manipuri)
Source Language (English)
Inflectional Intermediate Language 1
Intermediate Language 2(Inflectional+ Derivational)
Target Language (Manipuri)

Figure 7: Two different models of English-Manipuri translation

For word-based language modeling, 200,000 Manipuri news sentences are used. The approach is to take account of the various function words by grouping after syntactic analysis of the source side sentences using Stanford Parser (de Marneffe and Manning, 2008) and map to the corresponding target side of Manipuri after morphological analysis. The baseline system is based on the surface word forms using the Moses (Koehn et al., 2007) default setting. A factored translation model can embed multiples of monolingual MT systems inside of it. So, the addition of 'derivational morphology' has similar effect that one additional intermediate language includes the factored form. Consider the sentence 'I should have done it' meaning 'ঐ মদু তৌরমগদবনি।',(*ei madu touramgadabanee*) the phrase 'should have done' meaning তৌরমগদবনি consist of the inflections as well as derivational morphemes. The meaning of 'should have' is the derivational morpheme i.e., রমগদব and নি is the copula. The inflectional model gives the output as 'ঐ মদু তৌনি' meaning 'I'll do it'. But, considering the inflectional + derivational model, the output is adequately addressed by the induction of the syntactic information from the source side. The mapping from the auxiliary verbs to the derivational morphemes can help to improve as an important factor.

The various models developed are evaluated using BLEU (Papineni et al., 2002) and (NIST Doddington, 2002) automatic scoring techniques. SRILM is used for language modeling (Stolcke, 2002).

1	**English:** The number of teachers required is 58 for arts subjects and 32 for science subjects. Reference: অৱাংপা ওজা মশিং আর্টস সবজেক্তনা ৫৮ অমসুং সাইন্স সবজেক্তনা ৩২ নি Baseline: অৱাংপা ওজা মশিং থাকি সাইন্সকি ৩২ আর্টসতা ৫৮ Inflectional: অৱাংপা ওজা মশিং থাকি সাইন্সকি ৩২ আর্টসতা ৫৮ নি Derivational+ Inflectional: সবজেকশিং অৱাংপা ওজা মশিং আর্টসনা ৫৮ অমসুং সাইন্সনা ৩২ নি	Most of the meaning is conveyed by all the models.
2	**English:** The branches of the tree spread out in all directions. Reference: উগী মশা ময়ামদুনা মায়কৈ থুদীংদা লোঙথোকই Baseline: হৌবা উ খন্দোকপা মায়কৈ Inflectional: উগী মশা লোঙথোকই Derivational + Inflectional: উগী মশাশিং মায়কৈ লোঙথোকই	Poor meaning is conveyed by all the models.
3	**English:** Termites do a lot of damage by eating the wood of buildings. Reference: লৈশৌনা য়ুমগী উ চাথোক্তুনা অমাঙঅতা কয়ামরুম থোকহল্লি Baseline: লৈশৌনা তৌবসি য়াম্লা মাঙহনবা মীওইশিংনা ময়ুমদগী উ Inflectional: লৈশৌনা মাঙহনবা মীওইশিংনা ময়ুম উ Derivational + Inflectional : লৈশৌনা য়ুমগী উ চাথোক্তুনা অমাঙবা থোকহল্লি	No meaning is conveyed by all the models except the output of Derivational + Inflectional model.
4	**English:** Contract works should be given only to those who would be able to carry out the work with sincerity and dedication. Reference: থবক নিংথিনা তৌগদবা মীদা ঠিকা পীগদবনি Baseline: কন্ট্রেক্ট ৱার্ক পীগদবনি ঙসিগী থুন্নাই অসিদা carrying হায়বসি শেংনা থবক গ্রুপশিংনা Inflectional: কন্ট্রেক্ট ৱার্ক শেংনা থবক পীগদবনি Derivational + Inflectional: কন্ট্রেক্ট ৱার্ক থবক গ্রুপশিংনা হায়বসি পীগদবনি	Poor meaning is conveyed by all the models.
5	**English:** The election office has reportedly intimidated the IFCD for taking up necessary measures. Reference: ইলেক্সন ডিপার্টমেন্টকি মায়কৈদগী আইএফসিডিদা দরকার লৈবা থবক পায়থত্নবা থঙহনখ্রে Baseline: ইলেক্সন ডিপার্টমেন্টকি মায়কৈদগী আইএফসিডিদা দরকার লৈবা থবক পায়থত্নবা থঙহনখ্রে Inflectional: ইলেক্সন ডিপার্টমেন্টকি মায়কৈদগী আইএফসিডিদা দরকার লৈবা থবক পায়থত্নবা থঙহনখ্রে Derivational + inflectional : ইলেক্সন ডিপার্টমেন্টকি মায়কৈদগী আইএফসিডিদা দরকার লৈবা থবক পায়থত্নবা থঙহনখ্রে	Full meaning is conveyed by all the models.

Figure 8: Output of various models

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

Table 1: Corpus Statistics

	BLEU	NIST
Baseline	13.045	4.25
Inflectional	15.237	4.79
Derivational+ Inflectional	15.824	4.85

Table 2: Automatic Evaluation Scores

Table 1 show the corpus statistics and table 2 shows the automatic evaluation scores. The incorporation of derivational morphemes improves the BLEU and NIST scores by capturing a larger coverage of word forms.

6 Conclusion and Future Direction

With the present work, we have identified a novel approach to integrate finer linguistic details into the translation model by taking into account of the syntactic information from the source side and derivational and inflectional morphemes from the target side. Though English-Manipuri parallel corpora is limited in size, the performance of the SMT system is improved by taking into account of the above mentioned morphemes and thus helping to address the data sparsity problem for developing SMT systems between English and highly agglutinative and morphologically rich language like Manipuri. Our stress is mainly on the noun and verb morphology. Figure 8 shows the variations in the output of different models based on subjective evaluation. The automatic evaluation metrics shows the improvement of the scores for translation models catering more linguistic morphemes than the baseline models. The scalability of the present task is to develop SMT system between English and morphologically rich languages but with limited parallel corpora.

Acknowledgments

I, sincerely, thank Dr. Zia Saquib, Executive Director, CDAC (Mumbai), Prof. Sivaji Bandyopadhyay, Jadavpur University, Kolkata and the two anonymous reviewers for their support and valuable comments.

References

Andreas Stolcke. 2002. SRILM – an extensible language modeling toolkit. In Proceedings of the Intl. Conf. on Spoken Language Processing.

Ankur Gandhe, Rashmi Gangadharaiah, Kartik Vishweswariah and Ananthkrishnan Ramanathan. 2011. Handling Verb Phrase Morphology in Highly Inflected Indian Languages for Machine Translation, In proceedings of the 5th International Joint Conference on Natural Language Processing, Pages 111-119, Chiang Mai, Thailand.

Ch. Yashawanta Singh. 2000. Manipuri Grammar. Rajesh Publications, New Delhi.

George Doddington. 2002. Automatic evaluation of Machine Translation quality using n-gram co-occurrence statistics. In Proceedings of HLT 2002, San Diego, CA.

Guido Minnen, John Carroll and Darren Pearce, 2001. Applied Morphological Processing of English, Natural Language Engineering, 7(3), pages 207-223

Kemal Oflazer, and I. Durgar El-Kahlout. 2007. Exploring Different Representational Units in English-to-Turkish Statistical Machine Translation, in Proc. of the 2nd Workshop on Statistical Machine Translation, pages 25–32.

Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In Proceedings of 40th ACL, Philadelphia, PA.

Kristina Toutanova, Hisami Suzuki and Achim Ruopp. 2008. Applying Morphology Generation Models to Machine Translation, in Proc. 46th Annual Meeting of the Association for Computational Linguistics.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. Stanford Typed Dependency Manual.

Nizar Habash. 2008. Four techniques for online handling of out-of-vocabulary words in Arabic-English statistical machine translation, In

Proceedings of Association for Computational Linguistics-08.

Ondřej Bojar. 2007. English-to-Czech FactoredMachine Translation. In Proc. of ACL Workshop on Statistical Machine Translation, pages 232–239, Prague.

Ondřej Bojar and Jan Hajič. 2008. Phrase-Based and Deep Syntactic English-to-Czech Statistical Machine Translation, Proceedings of the Third Workshop on Statistical Machine Translation, pages 143–146, Columbus, Ohio, USA.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, John D. Lafferty and Robert L. Mercer, 1992. Analysis, Statistical Transfer, and Synthesis in Machine Translation. In Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, pages 83-100, Montreal, Canada.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra and Robert L. Mercer, 1993. Mathematics of Statistical Machine Translation: Parameter Estimation, Computational Linguistics, pages 163-311.

Purna C. Thoudam. 1982. Nouns in Meiteiron, Linguistics of the Tibeto Burman Area, Vol 6.2, Spring 1982. http://sealang.net/sala/archives/pdf8/thoudam1981nouns.pdf.

Philipp Koehn and Hieu Hoang. 2007. Factored Translation Models, Conference on Empirical Methods in Natural Language Processing (EMNLP), Prague, Czech Republic.

Philipp Koehn, 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, Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic.

Reyyan Yeniterzi and Kemal Oflazer. 2010. Syntax-to-Morphology Mapping in Factored Phrase-Based Statistical Machine Translation from English to Turkish, In proceeding of the 48th Annual Meeting of the Association of Computational Linguistics, Pages 454-464, Uppsala, Sweden.

Shobhana L. Chelliah. 1997. A Grammar of Meithei. Mouton de Gruyter, Berlin, pages 77-92.

Thoudam Doren Singh and Sivaji Bandyopadhyay. 2006. Word Class and Sentence Type Identification in Manipuri Morphological Analyzer, Proceeding of MSPIL 2006, IIT Bombay, pages 11-17, Mumbai, India.

Thoudam Doren Singh and Sivaji Bandyopadhyay. 2008. Morphology Driven Manipuri POS Tagger, In proceedings IJCNLP-08 Workshop on NLPLPL, pages 91-98, Hyderabad, India.

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

Thoudam Doren Singh and Sivaji Bandyopadhyay. 2010b. Manipuri-English Example Based Machine Translation System, International Journal of Computational Linguistics and Applications (IJCLA), ISSN 0976-0962, pages 147-158

Thoudam Doren Singh and Sivaji Bandyopadhyay. 2010c. 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.

Thoudam Doren Singh and Sivaji Bandyopadhyay. 2011a. Bidirectional Statistical Machine Translation of Manipuri English Language Pair using Morpho-Syntactic and Dependency Relations, In International Journal of Translation, Vol. 23, No.1 (Jan-Jun), 2011, pages 115-137.

Thoudam Doren Singh and Sivaji Bandyopadhyay. 2011b. Integration of Reduplicated Multiword Expressions and Named Entities in a Phrase Based Statistical Machine Translation System, Proceedings of the 5th IJCNLP, pages 1304–1312, Chiang Mai, Thailand.