Building Parallel Corpora for SMT System: A Case Study of English-Manipuri

Thoudam Doren Singh
Centre for Development of Advanced Computing
Gulmohor Cross Road No 9
Juhu, Mumbai-400049, INDIA

ABSTRACT
The Statistical Machine Translation (SMT) systems are developed using sentence aligned parallel corpus. The difficulty is that there is no parallel corpus at the required measure for many language pairs. The preparation of large scale parallel corpus takes time and demands the linguistics skill. In the present work, the various issues of a quality parallel corpus and a technique that extracts parallel corpus between Manipuri, a morphologically rich and resource constrained Indian language and English has been developed from a web based comparable news corpora. We explore the crux of the parallel corpora towards improving the translation quality through linguistics factors for the language pair.

General Terms
Parallel Corpus, SMT, Comparable Corpora

Keywords
Sentence alignment, Precision, Recall, English-Manipuri, Agglutinative, Morphology

1. INTRODUCTION
The data driven machine translation (MT) is the talk of the present MT research community. More importantly, a parallel corpus as a training data for Statistical Machine Translation (SMT) system is an essential requirement. In the present investigation, we try to identify what kind of parallel corpora is suitable for high performance SMT system between English and Manipuri. Manipuri is a morphologically rich and highly agglutinative in nature (Singh and Bandyopadhyay, 2006), (Singh and Bandyopadhyay, 2008). 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. Language resource for this language pair is not available at the required measure. Tone is very prominent in Manipuri language. While Manipuri is a tonal language, a special treatment of these tonal words is absolutely necessary. SMT systems between English and morphologically rich, highly agglutinative languages suffer badly if the adequate training and language resource is not available and linguistics information of the individual morpheme is not taken into account. The present task focuses on the language resource harnessing from news corpus available in the web between the two languages for SMT system development.

2. PARALLEL CORPORA OVERVIEW
Focusing on the machine translation research, one of the most important aspects is to be addressed i.e., building a good parallel corpus at the outset. There are some of the hard questions to be answered on building a good quality parallel corpus. The questions are: will it be possible to have an optimized parallel corpus which we can say is the best for a machine translation system for a particular language pair? What are the essential parameters or components in the parallel corpora? What'll be the optimum size of the parallel corpus? And again, are there good techniques available to build gold standard parallel corpora? Despite all the efforts, the moot question are what kind of training data (comparable or parallel corpora) will be suitable for a good quality translation output. Though there are reports to predict on the possible size of the training data (Kolachina et al., 2012) using learning curves, it is yet to come to a conclusion.

The first automatic parallel text alignment attempt was made by (Gale and Church, 1991), which is based on the idea that long sentences will be translated into long sentences and short sentences into short ones. Their approach works remarkably well on language pairs with high length correlation, such as French and English. Alignment performance degrades when the length correlation breaks down, such as in the case of Chinese and English (Ma, 2006). Even the Gale-Church algorithm may fail at regions that contain several sentences with similar lengths for language pairs with high length correlation.

From the practical point of view, creating small parallel corpora manually could be relatively easier, but building a large one is hard and time consuming. On top of that, maintaining the quality of the translation quality between the language pair is much harder. In fact, verification on the quality of the bi-text is one of the toughest tasks. The implication of building quality parallel corpora between a language pair is a pre-requisite so as to build an MT system between the language pair.

More often, parallel corpora between the resource rich language pair is available than the resource poor languages. In other words, it is very common that there can be parallel corpora between two resource rich languages. Some of the large parallel corpuses are Europarl available at http://www.statmt.org/europarl/. The Canadian parliamentary proceedings (also known as Hansards) in English and French are large bitext available. However, it is very uncommon to avail good quality parallel corpora between a resource poor or less privileged language and a resource rich language. This has resulted in the unavailability of the good MT systems between the resource-poor and resource-rich which heavily depend on the training corpus.

Parallel and comparable corpora are used primarily for translation and contrastive studies (McEnery and Xiao, 2007). Nonetheless, comparable corpora are a useful resource for contrastive studies and translation studies when used in combination with parallel corpora. However, the comparable corpora can be a poor basis for contrastive studies if the

sampling frames for the comparable corpora are not fully comparable. The non-parallel and yet comparable corpora overcome the two limitations of parallel corpora, since sources for original, monolingual texts are much more abundant than translated texts. However, mining translations in comparable corpora is much more challenging than in parallel corpora. What constitutes a good comparable corpus, for a given task or per se, also requires specific attention: while the definition of a parallel corpus is fairly straightforward.

One of the problems is that most of the potential parallel texts on the comparable corpora, even if they prove to have parallel fragments, often contain non-parallel fragments as well, especially at the beginning or at the end. The parallel fragment can be located anywhere in the document pair. The parallel fragments begin and end anywhere in the text and also it is possible to skip one or several sentences without breaking the fragment.

For the terminology extraction, specialized parallel and comparable corpora are clearly of use while for the contrast of general linguistic features such as tense and aspect, balanced corpora are supposed to be more representative of any given language in general. Specialized parallel corpora can be especially useful in domain-specific translation research. For a comparable corpus, a sampling frame is essential. But for parallel corpora, sampling frame is irrelevant, because each of the corpus components is exact translations of each other. While most of the existing comparable corpora are also specialized, it is relatively easier to find comparable text types in different languages. Therefore, in relation to parallel corpora, it is more likely for comparable corpora to be designed as general balanced corpora.

In addition to providing assistance to human translators, parallel corpora constitute a unique resource for the development of MT systems. Parallel corpora have been used to develop computer-assisted translation (CAT) tools for human translators, such as translation memories (TM), bilingual concordances and translator oriented word processors.

3. DEVELOPING ENGLISH-MANIPURI PARALLEL CORPORA

Parallel corpus between the Indian language and English is not available in the required measure. One problem that arises with the use of one-to-one parallel corpus (i.e. containing only one version of translation in the target language) is that the translation only represents one individual's introspection, albeit contextually and contextually informed (Malmkjaer, 1998). One possible way to overcome this problem is to include as many versions of a translation of the same source text as possible. In other words, a source sentence can be translated into multiple possibly likely target sentences. Or in the other sense, a source sentence can be interpreted in multiple possibly likely target sentences. In the present task, we attempt to build parallel corpus between English and Manipuri. Manipuri uses Bengali script to represent the written text in the present study. Consider the following English to Manipuri translations examples by two bilingual translators.

1.

Boundaries of the nine districts of Manipur are clear and distinct but the issue of dual voters demands a serious analysis.

First translation:

মনিপুরগী ডিষ্ট্রিক্ট মাপনগী ঙমখৈশিং ময়েক শেংনা লৈরি অদুবু দ্যুয়েল ভোটরগী লাক্লিবা ৱাফমগী মতাংদা য়াম্না কুপ্না খন্নবা দরকার ওই ।

Second Translation:

মনিপুরগী ডিষ্ট্রিক্ট মাপনগী ঙমখৈশিং ময়েক শেংনা লৈরি অদুবু দ্যুয়েল ভোটরগী ৱাফমগী মতাংদা য়াম্না কুপ্না খন্নবা দরকার তাই ।

2.

The association also appealed to the people for suggestions for bringing about a radical change in the structure and functional organization of the said colleges.

First translation:

এসোশিয়েসন অদুনা প্রজা মিয়ামদা হায়রিবা কোলেজশিং অদুগী স্ট্রকচর অমদি ফংশনেল ওর্গেনাইজেসন্দা অচৌবা অহোংবা অমা পুরক্লবা খন্নবা য়াবা মোত্ পুথোরক্লবা অপিল তৌখি ।

Second Translation:

প্রজাশিংদা এসোশিয়েসন্দুনা হায়রিবা কোলেজশিংদুগী স্ট্রকচর অমদি ফংশনেল ওর্গেনাইজেসন্দা অচৌবা অহোংবা অমা পুরক্লবা খন্লনবা মোত্ পুথোরক্লবা অপিল তৌখি ।

Considering the first example and their corresponding two translations, we can see the no syntactic variations in the structure but word alternatives appear i.e., দ্যুয়েল ভোটরগী লাক্লিবা ৱাফমগী becomes দ্যুয়েল ভোটরগী ৱাফমগী and দরকার ওই becomes দরকার তাই in the second translation. However, looking at the second example and their two translations, we see both syntactic variations as well as word variations, i.e., এসোশিয়েসন অদুনা প্রজা মিয়ামদা becomes প্রজাশিংদা এসোশিয়েসন্দুনা and খন্নবা য়াবা becomes খন্লনবা in the second translation.

While this solution is certainly of benefit to translation studies, it makes the task of building parallel corpora much more difficult. It also reduces the range of data one may include in a parallel corpus, as many translated texts are translated once only. It is typically texts such as literary works where multiple translations of the same work are available. These works tend to be non-contemporary and the different versions of translations are usually spaced decades apart, thus making the comparison of these versions less meaningful. The effect of source language on the translations is strong enough for source data perceptibly different from the target language. As such, a uni-directional parallel corpus is a poor basis for cross-linguistic contrast. In this sense, well matched bi-directional parallel corpora can become the bridge that brings translation and contrastive studies together. It is difficult to generate possible hypotheses as to translations.

Building good parallel corpora is hard as building a good MT system of a language pair. Both are constrained by the fact on how the sentences are interpreted and how they are decoded. Though, there has been report on building parallel corpora, defining on how difficult is yet to be measured on the part of the correctness. At the same time, it is again difficult to judge on the how the sentences are translated by MT systems for a given language pair.

Even when we integrate multiple translation of a source sentence into an MT system, the translation quality is hardly improved. So, we can assume that the MT systems also suffer from this drawback of inability to tackle which phrase /chunk will be picked up at what time from one of the possible source. (Razmara et al., 2012) reports on the ensemble decoding outperforming various strong baselines including mixture model as an experiment on domain adaptation.

On the part of building parallel corpora two common approaches are (a) human translation (b) extraction from an existing comparable corpus either from web or from other digital form.

The human translation can be verified by a second translator. However, there is no end on the verification. Since, five different human translators may translate a source sentence in five different ways. Be it ambiguous sentence, or be it pragmatic, or it requires discourse resolution, the meaning of the source sentence should be fully conveyed to the target sentence.

The first Manipuri-English parallel corpora development is reported by (Singh and Bandyopadhyay, 2010) using a semi-automatic approach from the comparable corpora collected from http://www.thesangaiexpress.com/. The web walked into the ACL meetings starting in 1999 as a source of linguistic data. In the recent times, there is more number of news published either in English or Manipuri online. Some of the websites are http://www.ifp.co.in/ published in English; http://www.poknapham.in/ published in Manipuri while http://www.hueiyenlanpao.com/ is published in both English and Manipuri. The current task is an extension of the semi-automatic Manipuri-English parallel corpora extraction approach which is based on the similarity based sentence alignment method using a bilingual lexicon by introducing morphological information after applying the Gale and Church sentence alignment process on the manually aligned paragraph. The bilingual lexicon which is used during the similarity based extraction is augmented with named entity list and transliterated list. One of the drawbacks of Manipuri language is that several language specific tools, such as morphological analyzer, POS tagger, named entity recognizer are not available in the required measure to help improving the parallel corpora extraction process. In the process, the sentence alignment accuracy is measured using precision and recall.

4. ENGLISH-MANIPURI SMT SYSTEM

The success story of the MT systems, in the recent times, more often SMT systems for the resource rich language pairs are reported with reasonable output. Accepting the fact that a prototype SMT systems can be developed for language pairs with decent translation quality provided the parallel corpora between the languages pair is available. Though some of the linguistic features can be incorporated into the SMT systems using factored translation models (Koehn and Hoang, 2007) based on Moses (Koehn et al., 2007), question still remains with local co-reference and more importantly ambiguity issues. Some of the important reason behind this is the data insufficiency and difficulty to integrate some of the language specific issues. 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), (Singh and Bandyopadhyay, 2010c), (Singh and Bandyopadhyay, 2011a) and (Singh and Bandyopadhyay, 2011b). This language contains abundant reduplicated multiword expressions (RMWE). The integration of these RMWE into the SMT is reported in (Singh and Bandyopadhyay, 2011b). In this process, the RMWE list is augmented to the training data. While Manipuri uses Bengali script to represent the text, the wide variations of tone are not captured during the textual representation. So, lexical ambiguity is very common in this language. This has resulted towards the requirement of a word sense disambiguation module.

In the present work, we filter the parallel corpus by considering few specific parameters. Manipuri is very rich in RMWE (re-duplicated multiword expression) (Singh and Bandyopadhyay, 2010). Integrating RMWE in the sentence alignment process by augmenting the RMWE list in the bilingual lexicon for the similarity based extraction could improve the precision and recall. The RMWE extraction and alignment is carried out based on the task of (Singh and Bandyopadhyay, 2010d) and (Singh and Bandyopadhyay, 2011b). Though building parallel corpora in a pure manual fashion requires sizable time slots and cost. By going through a set of experiments on English-Manipuri parallel corpus, we have sorted out the important parameters that can help to improve the translation quality by a reasonable BLEU and NIST scores. Finding an optimum size of parallel corpus for developing SMT systems for a given language pair is itself an important issue. While Manipuri is a highly agglutinative and morphologically rich language, the importance of handling each and every morpheme at the very basic level is necessary by proper translation representation at the corresponding sentence level. For the Indian languages, (Gandhe et al., 2011) used a technique to enrich the verb phrase morphology for English-Hindi translation systems using a set of templates to address a bit of data sparsity issue. 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.

4.1 Experimental Setup

The present English-Manipuri SMT is a phrase based system. In the experimental setup, Moses (Koehn et al., 2007) is used along with Giza++ (Och and Ney, 2003) alignment tool. SRILM toolkit for language modeling to develop target language model using 2,00,000 Manipuri news sentences from news corpus. BLEU (Papineni, 2002) and NIST (Doddinton, 2002) are used for automatic scoring. Automatic metrics and subjective evaluation are widely used on the part of MT system evaluation. In some cases, using statistical significance testing is found to be helpful. Tuning using MERT (Och, 2003) setting the mata-parameters such as maximum phrase length, maximum allowed distortion at their default values is carried out. Our experiment is carried out under the controlled environment. The sentence lengths of the training parallel sentences are set 40, 50, 60, 70 and 80 words. Table 1 shows the statistics of the corpus used in the experiment.

Table 1: Corpus Statistics

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

Table 2 shows the output of different lengths of the given input sentence of the SMT system.

Input Sentence: one of the gates at samushang is meant for draining out the flood water from lamphelpat area.

Reference Sentence: শামুশঙদা লৈরিবা গেটশিংগী মনুংদা অমদি লম্ফেলপাত্কি ঈচাওগী ঈশিং চিংথোক্লবা থম্মী.

Table 2: Test Output of different length

Length	Output
Default	অমা য়াই চিংথোক্লবা থম্মী লম্ফেলপাত্কি ঈশিং মী মতমদা ঈশিং ইমায় ৱাংখল্লক্লবদি গেট খিংলগা অমসুং অসি ইঞ্জিননা ওইনা চিংথোক্লগা .
40 words	অমা শামুশঙদা ডিপার্টমেন্ট লম্ফেলপাত্কি ঈশিং চিংথোক্লবা থম্মী.
50 words	অমা থম্মী শামুশঙদা ডিপার্টমেন্ট অসিগী গেট লম্ফেলপাত্কি ঈশিং চিংথোক্লবা.
60 words	অমা গেটতা লৈবা শমুশঙ অসি য়ুনিয়ন ফোর লোয়নদুনা ঈশিং ফাওবা ওফিস.
70 words	অমা শামুশঙদা draining লম্ফেলপাত্কি ঈশিং অমদগী মরমদি.
80 words	অমা শামুশঙদা চিংথোক্লবা ডিপার্টমেন্ট থম্মী অসিগী গেট ঈশিং.

The outputs of shorter sentences length training corpus are closer to the reference translation as compared to the longer sentences. The Table 3 shows the automatic BLEU and NIST scores using the named entities and reduplicated multiword expression in the bilingual lexicon during the parallel corpora extraction technique and Table 4 shows the automatic scores of based on the sentence length.

Table 3: Automatic Evaluation using NE and RMWE in bilingual lexicon

	BLEU	NIST
Baseline (default length)	13.045	4.25
NE in bilingual lexicon	14.237	4.59
NE + RMWE in bilingual lexicon	14.824	4.75

Table 4: Automatic Evaluation based on Sentence length

	BLEU	NIST
Baseline (default length)	13.045	4.25
Length <15	13.567	4.60
Length <40	13.637	4.79
Length <50	13.824	4.85
Length <60	13.941	4.90
Length <70	13.842	4.81
Length <80	13.341	4.70

5. CONCLUSION

From the current the task, what is lacking at the moment is a full automatic technique to extract parallel corpora from a comparable corpus for developing SMT systems. Over and above, the verification of the quality of the aligned sentences should be carried out prior to feeding into the SMT systems. A suitable pre-processing technique based on the word order of the source and target language can be applied. We found that setting an appropriate sentence length of the training parallel corpus in this language pair is definitely helpful. In future, various experiments can be conducted in order to decide whether we are going in the right direction towards the development of SMT system with all the possible techniques at the preprocessing, intermediate and post processing stages and also based on the word order of the source and target sentence along with exploitation morphological information. On experiencing, building parallel corpora and MT system, there is a high inter-correlation between the two interdependent activities.

6. ACKNOWLEDGMENTS

I, sincerely, thank Dr. Zia Saquib, Executive Director, CDAC Mumbai for his support and Professor Sivaji Bandyopadhyay, Jadavpur University, Kolkata for his valuable input.

7. REFERENCES

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

[2] Gale, W. A., Church, K. W., 1991. A program for aligning sentences in bilingual corpora, In proceedings of 29th Annual meeting of ACL, Pages 177-184, Berkeley, California

[3] Gandhe, A., Gangadharaiah, R., Vishweswariah K., Ramanathan, A. 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, 2011.

[4] Koehn, P., Hoang, H. 2007. Factored Translation Models, Conference on Empirical Methods in Natural Language Processing (EMNLP), Prague, Czech Republic.

[5] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E. 2007. Moses: Open Source Toolkit for Statistical Machine Translation, Annual Meeting of the Association for Computational Linguistics (ACL), Demonstration session, Prague, Czech Republic.

[6] Kolachina, P., Concedda, N., Dymetman, M., Venkatapathy, S., 2012. Prediction of learning curves in Machine Translation, In proceeding of the 50th Annual meeting of the ACL, Pages 22-30, Jeju, Korea.

[7] Ma, X., 2006. Champollion: A Robust Parallel Text Sentence Aligner. In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC). Genova, Italy.

[8] Malmkjaer, K., 1998. Ed. Translation and Language Teaching: Language Teaching and Translation, Manchester, UK.

[9] McEnery, A., Xiao, Z. 2007. Parallel and comparable corpora? In Incorporating Corpora: Translation and the Linguist. Translating Europe. Multilingual Matters, Clevedon, UK.

[10] Och, F. J., Ney, H. 2003. A Systematic Comparison of Various Statistical Alignment Models, Computational Linguistics, volume 29, number 1, Pages. 19-51.

[11] Och, F. J., 2003. Minimum error rate training in Statistical Machine Translation, In the proceeding of Proceedings of ACL.

[12] Papineni, K., Roukos, S., Ward, T., and Zhu, W. 2002. BLEU: a method for automatic evaluation of machine translation. In Proceedings of 40th ACL, Philadelphia, PA.

[13] Razmara , M., Foster, G., Sankaran, B., Sarkar, A. 2012. Mixing multiple translation models in Statistical Machine Translation, In proceedings of the 50th Annual Meeting of the Association of Computational Linguistics (ACL 2012), Juju Island, Korea.

[14] Singh, T. D., Bandyopadhyay, S. 2006. Word Class and Sentence Type Identification in Manipuri Morphological Analyzer, Proceeding of MSPIL 2006, IIT Bombay, Pages 11-17, Mumbai, India.

[15] Singh, T. D., Bandyopadhyay, S. 2008. Morphology Driven Manipuri POS Tagger, In proceedings of IJCNLP-08 Workshop on NLPLPL, Pages 91-98, Hyderabad, India.

[16] Singh, T. D., Bandyopadhyay, S. 2010a. 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.

[17] Singh, T. D., Bandyopadhyay, S., 2010b, Manipuri-English Example Based Machine Translation System, International Journal of Computational Linguistics and Applications (IJCLA), ISSN 0976-0962, Pages 147-158

[18] Singh, T. D., Bandyopadhyay, S. 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.

[19] Singh, T. D., Bandyopadhyay, S. 2010d. 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.

[20] Singh, T. D., Bandyopadhyay, S. 2011a, Bidirectional Statistical Machine Translation of Manipuri English Language Pair using Morpho-Syntactic and Dependency Relations, In International Journal of Translation (IJT), ISSN 0970-9819, Vol. 23, No.1 (Jan-Jun), 2011, Pages 115-137.

[21] Singh, T. D., Bandyopadhyay, S. 2011b, Integration of Reduplicated Multiword Expressions and Named Entities in a Phrase Based Statistical Machine Translation System, Proceedings of the 5th International Joint Conference on Natural Language Processing, Pages 1304–1312, Chiang Mai, Thailand, November 8 – 13, 2011.

[22] Stolcke, A. 2002. SRILM – an extensible language modeling toolkit. In Proceedings of the International Conference on Spoken Language Processing.