Evaluation of English-Myanmar Syntactic Reordering System by Using Reordering Metrics

Thinn Thinn Wai, Tin Myat Htwe, Ni Lar Thein
Univesity of Computer Studies, Yangon
thin2wai@gmail.com, tinmyathtwe@gmail.com,nilarthein@gmail.com

Abstract

Because different languages employ different word order in their syntax, one requirement of machine translation (MT) system is to get the target word in the right order. While phrase-based machine translation system do very well at short-distance reordering, long-distance reordering seems to be a challenging task. One way of overcoming this challenge is to use linguistic information and reorder the input sentence so that the word order is consistent with what the target language might expect. In this paper, long-distance reordering is solved by using the linguistic information such as part-of –speech tags and syntactic function tags. Moreover, the quality of word order is also measured by using reordering metrics. This reordering system is implemented for using as a component in English-Myanmar Translation system.

1. Introduction

In statistical machine translation, the use of reordering strategies allows for an important improvement in translation accuracy. The first SMT systems introducing word reordering were based on the brute-force approach, as (in principle) they applied all permutations of the input words in order to find out the one producing the right target word order. When a English sentence is translated into Myanmar sentence, the verb in the English sentence must be moved towards the end of the English sentence in order to obtain the correct Myanmar word order. On a sub sentential level, Myanmar word order diverges from English mostly within the noun phrase and verb phrase. Without reordering, the particles can be far from their relative nouns, verbs and adjectives and the correct word order can’t be obtained. In addition to this, the meaningful translation can’t also be obtained. Therefore, reordering is necessary for translation from English language to Myanmar Language. In this work, long-distance reordering for English-Myanmar machine translation and evaluation of implemented reordering system are presented.

2. Related Work

Different approaches have been developed to deal with the word order problem. First approaches worked by constraining reordering at decoding time [7]. In [12], the alignment model introduced the restrictions in word order, which leads also to restrictions at decoding time. A comparison of these two approaches can be found in [2]. They have in common that they do not use any syntactic or lexical information; therefore they rely on a strong language model or on long phrases to get the right word order. Other approaches were introduced that use more linguistic knowledge, for example the use of bitext grammars that allow parsing the source and target language [13]. In [10], syntactic information was used to re rank the output of a translation system with the idea of accounting for different reordering at this stage. In [11], a lexicalized block-oriented reordering model is proposed that decides for a given phrase whether the next phrase should be oriented to its left or right.

The most recent and very promising approaches that have been demonstrated reorder the source sentences based on rules learned from an aligned training corpus with a POS-tagged source side [8, 9, 20]. These rules are then used to reorder the word sequence in the most likely way.

In our approach we follow the idea proposed in [20] of using a parallel training corpus with a tagged source side to extract rules which allow a reordering before the translation task. Moreover, translation models are still not close to modeling the reordering performance of human translators; and reordering is an important predictor of the quality of translation output. In this paper, novel metrics of reordering which directly measure word order differences between human reference sentences and machine translations are used to evaluate the reordering performance.

3. The differences of syntax structure between English and Myanmar

English is SVO language and Myanmar is SOV language. While SOV structure is used in formal translation, OSV structure is mostly used in informal (colloquial) translation. Generally, Myanmar is a verb final language. Significant differences of word order in these languages are mostly found in adjective, adverb, preposition and modal auxiliary verb. In clause level, the dependent clause in English always translated before independent clause translation in Myanmar. Therefore, both of local reordering (short-distance) and global reordering (long-distance) are needed to translate English to Myanmar or Myanmar to English. These two reordering can be seen in Figure 1.

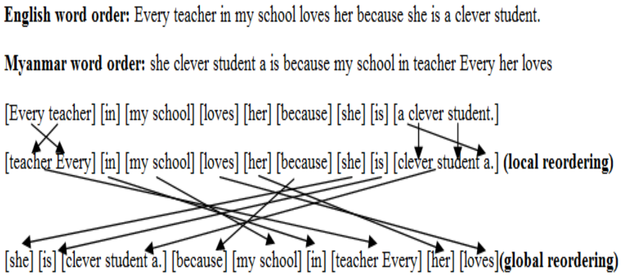

English word order: Every teacher in my school loves her because she is a clever student.
Myanmar word order: she clever student a is because my school in teacher Every her loves
[Every teacher] [in] [my school] [loves] [her] [because] [she] [is] [a clever student.]
[teacher Every] [in] [my school] [loves] [her] [because] [she] [is] [clever student a.] **(local reordering)**
[she] [is] [clever student a.] [because] [my school] [in] [teacher Every] [her] [loves]**(global reordering)**

Figure 1. Local and global reordering in English-Myanmar Reordering

4. System Overview

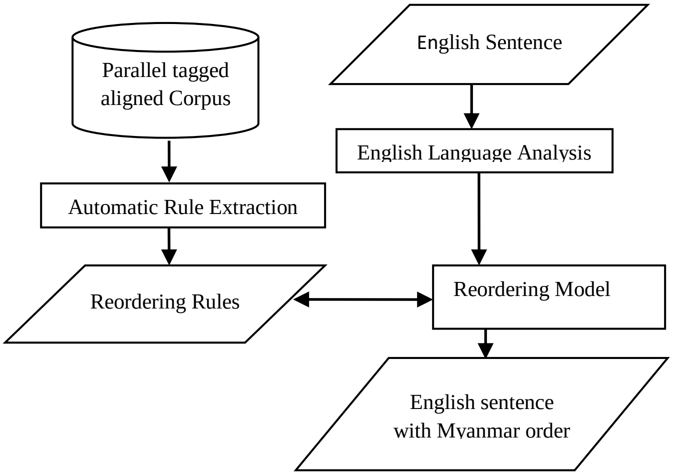

Parallel tagged aligned Corpus
Automatic Rule Extraction
Reordering Rules
English Sentence
English Language Analysis
Reordering Model
English sentence with Myanmar order

Figure 2. Overview of the system

As shown in Figure 2, there are two key components in this reordering system; rule generation and reordering. In rule generation, English-Myanmar reordering rules are automatically extracted from tagged-aligned corpus by using rule extraction procedure described in [27].

Therefore, Tagged-aligned corpus is created firstly by aligning and analyzing the plain text corpus . This corpus is created over 1000 sentences, 1000 compound sentences and 500 complex sentences. In reordering component, the input sentence is firstly analyzed by using English Language Analyzer to extract syntactic structure. By using the pos sequences and function tag sequences obtained from analysis phase, reordering is then performed. In this reordering model local reordering, short-distance reordering is solved by part-of-speech reordering rules and global reordering; long-distance reordering is solved by function-tag reordering rules.

5. Markov process reordering model

In this reordering model we proposed, function tag and pos tag sequence are taken as input. And then, reordering is performed based on First Order Markov model by using the alignment probabilities extracted from corresponding function tag and pos tag reordering rules. The tag alignment sequence a_1^K specifies a reordering of source tag sequences into target language tag order. In this way the source language tag sequence u_1^K is reordered into $u_{a_1}, u_{a_2}, \ldots, u_{a_k}$ under the model $P(a_1^k \backslash u_1^k, K, e_1^k)$. A First Order Markov process is firstly defined over tag alignment sequences $a_k \in \{1,2,...,K\}$ as can be seen in (1) and these alignment sequences are obtained from the extracted reordering rules from tagged aligned corpus.

$$P(a_1^k \backslash u_1^k, K, e_1^k) = P(a_1^k \backslash u_1^k)$$

$$= P(a_1) \prod_{k=2}^{K} P(a_k \backslash a_{k-1}, u_1^K) \qquad (1)$$

This reordering model involves two acceptors un-weighted permutation acceptor U and weighted permutation acceptor H. Un-weighted permutation acceptor U contains all reordering of the source language tag sequence. The second acceptor H assigns alignment probabilities to a given reordering a_1^k of the source tag sequence u_1^k. According to the maximum alignment probability, optimal reordering rule is extracted and words are reordered by optimal reordering rule. This reordering model performs in hierarchical level; word level and chunk level. For word level reordering, it uses the part-of-speech tag alignment probabilities and function tag alignment probabilities are used for chunk-level-reordering.

6. Importance of Reordering Metrics

Machine translation research relies heavily upon automatic metrics to evaluate the performance of models. However, current metrics rely upon indirect methods for measuring the quality of the word order, and their ability to capture reordering performance has been demonstrated to be poor .

Traditionally, there are two main approaches to capture reordering. The first way to measure the quality of word order is to count the number of matching n-grams between the reference and the hypothesis. This is the approach taken by the BLEU score. This method discounts any n-gram which is not identical to a reference n-gram, and also does not consider the relative position of the strings. They can be anywhere in the sentence. Another common approach is typified by METEOR and TER. They calculate an ordering penalty for a hypothesis based on the minimum number of chunks the translation needs to be broken into in order to align it to the reference. The disadvantage of the second approach is that aligning sentences with very different words can be inaccurate. Also there is no notion of how far these blocks are out of order. More sophisticated metrics, such as the RTE metric (Pad´ o et al., 2009), use higher level syntactic or even semantic analysis to determine the quality of the translation.

Therefore, Brich et.al proposed Lexical Reordering Metric (LRscore) which is combined the permutation distance metric with lexical metric to measure the quality of word order. LRscore is language independent. The reordering component relies on abstract alignments and word positions and not on words at all. The lexical component of the system can be any meaningful metric for a particular target language. In this work, Brich et.al idea is followed to measure the quality of word order of the proposed reordering system. In this experiment 3-gram BLEU is used for lexical metric. To perform this experiment, alignment metrics are changed to permutation distance metrics firstly. Secondly, Hamming distance and Kendall's Tau distances are calculated over the system generated translation permutation and reference permutation. Thirdly, lexical metric (BLEU score) is calculated for the translation hypothesis. Finally LRscore is calculated by combining the permutation distances metrics and lexical metric (BLEU). These works are detail explained in the following sections.

7. Changing alignment metrics into permutation metrics.

The relative ordering of words in the source and target sentences is encoded in alignments. So alignments are interpreted as permutations and permutations are used to evaluate reordering performance. The ordering of the words in the target sentence can be seen as a permutation of the words in the source sentence. The source sentence s of length N consists of the word positions $s_0,s_1,s_2,\ldots,s_n$. Using an alignment function where a source word at position i is mapped to a target word at position j with the function $a : i \rightarrow j$ the source word positions can be reordered to reflect the order of the words in the target. This gives us a permutation.

A permutation is a bijective function from a set of natural numbers 1, 2, · · · ,N to itself. These permutation are named as π and σ. Permutations encode one-one relations, whereas alignments contain null alignments and one-many, many-one and many-many relations. Source words aligned to null $(a(i) \rightarrow null)$ are the target word position immediately after the target word position of the previous source word $(\pi(i) = \pi(i-1+1))$. Where multiple source words are aligned to the same target word or phrase, a many-to-one relation, the target ordering is assumed to be monotone. When one source word is aligned to multiple target words, a one-to-many relation, the source word is assumed to be aligned to the first target word. Alignment and permutation extraction are shown in Figure 3.

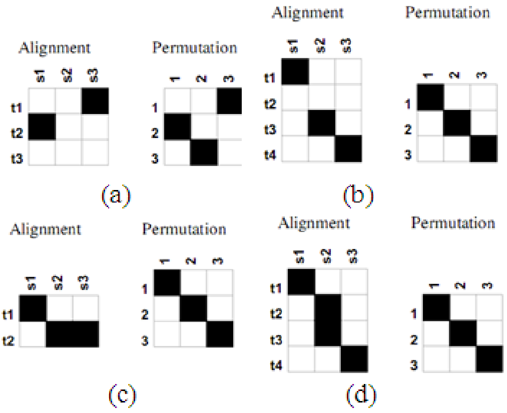


Figure 3 . Alignments and permutations containing (a) null alignment (b) one-to-one alignment (c) many-to-one alignment and(d) one-to-many alignment.

8. Permutation Distance Metrics

Here, the calculation of the two permutation metrics Hamming Distance and Kendall’s Tau Distance metric are explained. All these permutation metrics are subtracted from 1 to scale these metrics in order to easily compare them with other metrics. Moreover, the lexical reordering metric (BLEU) and combined metric (LRscore) is also explained.

8.1. Hamming Distance

The Hamming distance (Hamming, 1950) measures the number of disagreements between two permutations. The Hamming distance for permutations was proposed by (Ronald, 1998) and is also known as the exact match distance. It is defined as follows;

$$d_h(\pi,\sigma)=1-\frac{\sum_{i=1}^{n} x_i}{n}, x_i=\{^{0 if\, \pi(i)=\sigma(i)}_{1 otherwise}$$

where n is the length of the permutation. The Hamming distance will calculate the percentage of words in the translation which are in exactly the same order as in the reference sentence.

8.2. Kendall’s Tau Distance

Kendall’s tau distance is the minimum number of transpositions of two adjacent symbols necessary to transform one permutation into another (Kendall, 1938; Kendall and Gibbons, 1990). Kendall’s tau seems particularly appropriate for measuring word order differences because it measures relative differences. It has been used as a means of estimating the distance between a system-generated and a human-generated gold standard order for the sentence ordering task. The number of transpositions can be calculated by counting the number of crossings. It is defined by the following equations.

$$d_k(\pi,\sigma)=1-\sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} x_{ij}}{z}}$$

,

$$x_{ij} = \begin{cases} 1 \text{ if } \pi(i) < \pi(j) and \sigma(i) > \sigma(,j) \\ 0 \text{ otherwise} \end{cases}$$

$$z=\frac{(n^2-n)}{2} \quad ,$$

n= the length of the permutation.

8.3. BLEU

BLEU (Bilingual Evaluation Understudy) is an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another. Quality is considered to be the correspondence between a machine's output and that of a human: "the closer a machine translation is to a professional human translation, the better it is".BLEU was one of the first metrics to achieve a high correlation with human judgements of quality,[2][3] and remains one of the most popular. Scores are calculated for individual translated segments—generally sentences—by comparing them with a set of good quality reference translations. BLEU uses a modified form of precision to compare a candidate translation against multiple reference translations. It is defined as follows:

$$BLEU = BP \bullet \exp(\sum_{n=1}^{N} w_n \log p_n)$$

where p_n is the modified n-gram precision, BP is the Brevity penalty. Moreover, N is defined over which language model is used. Here, tri-gram language model is used and so N=3 and $w_n = 1/N$ =1/3.

8.4. Combined Metric

The LRscore is a linear interpolation of a reordering metric with the BLEU score. It consists of a reordering distance metric which is linearly interpolated with a lexical score to form a complete machine translation evaluation metric. The metric is decomposable because the individual lexical and reordering components can be looked at individually. The following formula describes how to calculate the LRscore:

$$LRscore = \alpha R + (1 - \alpha) L$$

The metric contains only one parameter, α, which balances the contribution of the reordering metric, R, and the lexical metric, L. Here we use BLEU as the lexical metric. R is the average permutation distance metric adjusted by the brevity penalty and it is calculated as follows:

$$R = \frac{\sum_{s \in S} d_s BP_s}{|S|}$$

where S is a set of test sentences, d_s is the reordering distance for a sentence and BP is the brevity penalty. The brevity penalty is calculated as:

$$BP = \begin{cases} 1 & \text{if t>r} \\ e^{1-r/t} & \text{if t} \leq \text{r} \end{cases}$$

where t is the length of the translation, and r is the closest reference length. If the reference sentence is slightly longer than the translation, then the brevity penalty will be a fraction somewhat smaller than 1.

9. Evaluation of reordering system with Reordering Metrics

Experiment is carried out by comparing the system output with both of formal and informal (colloquial) reference of translation. To evaluate the performance of reordering system, combined metric (LRscore) is used. The overall performance of this reordering system is measured on 1000 tested sentence. Average percentage of Hamming Distance and Kendall's Tau can be seen in Table 4. Moreover, the percentage of LRscore of 3-gram BLEU and unigram BLEU on Hamming distance calculation (LR-HB1 and LR-HB3) and the percentage of LRscore of 3-gram BLEU and unigram BLEU on Kendall's Tau distance calculation(LR-KB1 and LR-KB3) are described in Table 5.

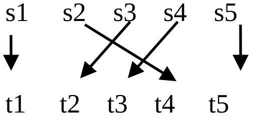
Table 1. Hamming Distance Calculation example for formal reference

Input Sentence	I like him very much because he is a good friend for me.
System Output	he(s1)me(s2)for(s3)friend good a (s4) is(s5)because(s6)I(s7)him(s8)like(s9). n=9
Colloquial Reference	သူသည်(t1)ကျွန်ုပ်(t2)အတွက်(t3)သူငယ်ချင်း ကောင်းတစ်ယောက်(t4)ဖြစ်(t5)သောကြောင့်(t6) ကျွန်ုပ်သည်(t7)သူ့ကို(t8)နှစ်သက်ပါသည်။(t9)
Hamming distance calculation	$d_h(\pi,\sigma) = 1 - \frac{\sum_{i=1}^{n} x_i}{n}$ $= 1 - \frac{0+0+0+0+0+0+0+0+0}{4}$ $= 1 - 0 = 1 = 100\%$

Table 2. Hamming Distance Calculation example for informal reference

Input Sentence	He goes to school.
System Output	He(s1)school(s2)to(s3)go(s4). n=4
Colloquial Reference	သူ(t1)ကျောင်း(t2) သွားသည်။(t3).
Hamming distance calculation	$d_h(\pi,\sigma)=1-\frac{\sum_{i=1}^{n}x_i}{n}$ $=1-\frac{0+0+1+1}{4}$ $=1-\frac{2}{4}=0.5=50\%$

Table 3. Kendall's Tau distance calculation example for informal reference

Input Sentence	He bought a shirt from the market.
System Output	He(s1)shirt a(s2)market(s3) from (s4)bought (s5). n=4
Colloquial Reference	သူ(t1)ဈေး (t2) မှ(t3) ရှပ်အင်္ကျီတစ်ထည် (t4) ဝယ်ခဲ့သည်။(t5)
Hamming distance calculation	 Number of crossing =3 $d_k(\pi,\sigma)=1-\sqrt{\frac{3}{\frac{25-5}{2}}}=\sqrt{\frac{3}{10}}$ $=1-0.55=0.45=45\%$

Tabel 4. Average percentage on 1000 tested sentence.

	Hamming Distance	Kendall's Tau Distance
Formal Reference	91.25%	79.51
Informal Reference	65.51%	75.61%

Table 5. Average Percentage of LRscore

	LR-HB1	LR-HB3	LR-KB1	LR-KB3
Average Percentage	68.5	70.1	72.5	74.1

10. Conclusions

In this work, short-distance and long-distance reordering for English-Myanmar translation are solved by using pos reordering rules and function tag reordering rules. Proposed reordering system works well in the sentences that the analyzer can tag the function tags correctly. The analyzer used in this proposed reordering system can work well in the sentences that have word length less than or equal to 20. Moreover, the performance of this reordering system is evaluated by using the permutation distance metrics and lexical reordering Metrics. By learning the experimental results of the permutation distance metrics , proposed reordering system can be seen that the performance of this reordering system falls in colloquial references while it have good performance in formal reference. Moreover, lexical reordering metrics (LRscore) can be found that it can be meaningful and accurately measures the word order performance of the translation model.

References

1] C. Tillmann and H. Ney. 2002. Word reordering and DP beam search for statistical machine translation to appear in Computational Linguistics.

[2] R. Zens and H. Ney. 2003. A comparative study on reordering constraints in statistical machine trans lation. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, vol ume 1, pages 144–151, Sapporo, Japan.

[3] S. Vogel, F.J. Och, C. Tillmann, S. Nießen, H. Sawaf, and H. Ney. 2000. Statistical methods for machine translation. InW.Wahlster, editor, Verbmobil: Foundations of Speech-to-Speech Translation, pages 377–393. Springer Verlag: Berlin, Heidelberg, New York.

[4] Y.Y. Wang and A. Waibel. 1997. Decoding algorithm in statistical translation. In Proc. 35th Annual Meeting of the Assoc. for Computational Linguistics, pages 366–372, Madrid, Spain, July.

[5] Ei Ei Han and Ni Lar Thein, "Morphological Synthesis For Myanmar Language", Proceeding of International Conference on Internet Information Retrieval, Korea, 2007.

[6] Yaser Al-Onaizan and Kishore Papineno. 2006. Distortion models for statistical machine translation. In Proceedings of the 21st International Conference on Computational Linguistics and the 4th annual meeting of the ACL, pages 529–536, Sydney, Australia.

[7] A. L. Berger, S. A. Della Pietra, and V. J. Della Pietra,1996. A maximum entropy approach to natural language processing. Computational Linguistics, 22(1):39.

[8] B. Chen, M. Cettolo, and M. Federico. 2006. Reordering rules for phrase-based statistical machine translation. In Int. Workshop on Spoken Language Translation Evaluation Campaign on Spoken Language Translation, pages 1–15.

[9] M. Popovic and H. Ney. 2006. POS-based word reorderings for statistical machine translation. In Proc. of the 5th Int. Conf. on Language Resources and Evaluation (LREC), page 1278, Genoa, Italy.

[10] L. Shen, A. Sarkar, and F. J. Och. 2004. Discriminative reranking for machine translation. In HLTNAACL 2004: Main Proc., page 177.

[11] C. Tillmann and T. Zhang. 2005. A localized prediction model for statistical machine translation. In Proceedings of the 43rd Annual Meeting of the Assoc. for Computational Linguistics (ACL), pages 557–564, Ann Arbor, MI.

[12] D. Wu. 1996. A polynomial-time algorithm for statistical machine translation. Proc. 34th Annual Meeting of the Assoc. for Computational Linguistics, page 152.

[13] D. Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. Computational Linguistics, 23(3):377.

[14] Y. Zhang, R. Zens, and H. Ney. 2007. Chunk-Level Reordering of Source Language Sentences with Automatically Learned Rules for Statistical Machine Translation. In Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL): Proceedings of the Workshop on Syntax and Structure in Statistical Translation (SSST), pages 1–8, Rochester, NY.

[15] Myat Thuzar Tun and Ni Lar Thein, " English Syntax Analyzer for English-to-Myanmar Machine Translation", In proceedings of the Fifth International Conference on Computer Application, Myanmar, February, 8-9,2007.

[16] Myat Thuzar Tun, Tin Myat Htwe and Ni Lar Thein, "EMTM: An Effective Language Translation Model", In proceedings of International Conference on Internet Information Retrieval, Korea, November 30, 2005.

[17] Shankar Kumar “Local Phrase Reordering Models for Statistical Machine Translation”, Center for Language and Speech Processing, Johns Hopkins University, 3400 North Charles Street, Baltimore, MD 21218, U.S.A.

[18] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer, “The Mathematics of Statistical Machine Translation: Parameter Estimation,” Computational Linguistics, vol. 19(2), pp. 263–312, 1993.

[19] Kenji Yamada and Kevin Knight. 2000. A Syntax based Statistical Translation Model. ACL 2000.

[20] Josep M. Crego and Jose B. Marino. 2006. Reordering Experiments for N-Gram-based SMT. In Spoken Language Technology Workshop, pages 242-245, Palm Beach, Aruba.

[21] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, “BLEU: a Method for Automatic Evaluation of Machine Translation”, Association for Computational Linguistics, 2002, pp. 311-318.

[22] Satanjeev Banerjee and Alon Lavie. 2005. Meteor: Anautomatic metric for MT evaluation with improved correlation with human judgments. In Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization.

[23]Alexandra Birch, Phil Blunsom, and Miles Osborne.2010. Metrics for MT Evaluation: Evaluating Reordering. Machine Translation (to appear).

[24]Thinn Thinn Wai, Tin Myat Htwe, Ni Lar Thein. “Markov-based Reordering Model for English-Myanmar Translation”,SICE 2011, Tokyo, Japan.

[25] Thinn Thinn Wai, Tin Myat Htwe, Ni Lar Thein. “Automatic Reordering Rule Generation and Application of Reordering Rules in Stochastic Reordering Model for English-Myanmar Machine Translation” IJCA,vol-27,No-8,2011.

[26]Alexandra Birch, Phil Blunsom, and Miles Osborne.2010. “ LRscore for Evaluating Lexical and Reordering Quality in MT”.

[27] Thinn Thinn Wai, Tin Myat Htwe, Ni Lar Thein. “Automatic Reordering Rule Generation for English-Myanmar Translation” ICCA-2011.