Reordering System Implementation for Myanmar-English-Myanmar Machine Translation

Thinn Thinn Wai
University of Computer Studies, Yangon
thin2wai@gmail.com

Ni Lar Thein
University of Computer Studies, Yangon
nilarthein@gmail.com

Abstract

Reordering is one of the most challenging and important problems in Statistical Machine Translation. Without reordering capabilities, sentences can be translated correctly only in the case when both languages implied in translation have a similar word order. When translating is between language pairs with high disparity in word order, word reordering is extremely desirable for translation accuracy improvement. Our Language, Myanmar is a verb final language and reordering is needed when our language is translated from other languages with different word orders. In this paper, hierarchical rule-based reordering approach is used. This work is intended to be incorporated into Myanmar-English-Myanmar machine translation. Proposed reordering system also serves as a pre-translation reordering system and gain the accuracy (94.23%) in Myanmar-English reordering and (90.99%) in English-Myanmar reordering.

Key Words: Reordering, Statistical Machine Translation, rule-based reordering

1. Introduction

Reordering is a major challenge in Statistical Machine Translation (SMT). Reordering involves permuting the relative word order from source sentence to translation in order to account for systematic differences between languages. Correct word order is important not only for the fluency of output; it also affects word choice and the overall quality of the translations. Therefore, many methods and approaches are proposed for solving word order differences. Most of the approaches can solve the short-distance reordering, but long-distance reordering is still a challenging task.

English language has Subject-Verb-Object structure and Myanmar language has Subject-Object-Verb structure. Sometimes, Myanmar language has Object-Subject-Verb structure in colloquial sentences. Generally, Myanmar language is verb final language. Therefore, one English sentence can be translated as two reordering patterns; such as Subject-Object-Verb and Object-Subject-Verb basically. Moreover, the syntactic structure differences between Myanmar and English languages exist not only in word level but also in phrase level. Therefore, not only local reordering but also global reordering is needed to solve in English-Myanmar and Myanmar-English translation. In order to solve the local reordering, Part-of-Speech reordering rules extracted from morphological analysis is applied. For global reordering, function tag reordering rules taken out from syntactic analysis is used.

The plan of this paper is as follows. In the next section, related works which use reordering approaches in a pre-processing step are reviewed. In section 3, the overview of proposed reordering system is described. The step by step processes of proposed reordering system are presented in the sub sections 3.1, 3.2, and 3.3 respectively. Section 4 describes the evaluation of proposed reordering system and this paper is concluded in section 5 and also discusses the future work of the proposed system in this section.

2. Related Work

Different approaches have been developed to deal with the word order problem. First approaches worked by constraining reordering at decoding time [1]. In [2], 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 [3]. In [4], 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 [5], 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 [6, 7, and 8]. These rules are then used to reorder the word sequence in the most likely way.

In our approach the idea proposed in [8] is followed and the tagged aligned corpus is used to extract rules which allow a reordering before the translation task.

3. Overview of Proposed System

There are three key components in the proposed reordering system. They are

1. Tagged Aligned Corpus Creation
2. Automatic Reordering Rule Generation
3. Reordering

Tagged aligned corpus creation algorithm is shown in section 3.1 and automatic reordering rule generation algorithm is described in section 3.2. Then, the working style of reordering is presented in section 3.3. The overview of proposed reordering system can be seen in Figure 1.

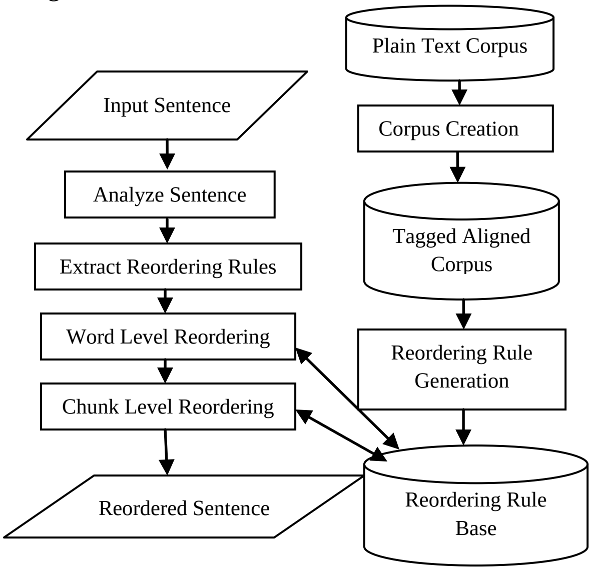

Figure1. Overview of Proposed System

3.1. Tagged Aligned Corpus Creation

In this work, two kind of tagged aligned corpora; Myanmar tagged aligned corpus and English tagged aligned corpus are created for using as the resources for automatic reordering rule generation algorithm. Tagged aligned corpus creation algorithm is shown in algorithm 3.1. In the corpus creation algorithm, sentences from plain text corpus are taken as the input and tagged aligned sentences are output and stored these tagged aligned sentences into the corpus.

Algorithm 3.1: Corpus Creation Algorithm

Input: Plain Sentences
Output: Tagged Aligned Sentences
begin
    load the plain text corpus
    extract each sentence from the corpus
    for each sentence
      begin
        analyzed sentence
        add alignment positions
        store tagged aligned sentence in the corpus
      end
end

For Myanmar tagged aligned corpus creation, Myanmar plain text corpus is used as a resource and English plain text corpus is applied for building English tagged aligned corpus. For analyzing the input Myanmar sentences, Myanmar Word Segmentor [9], bi-gram POS tagger [10] and Naïve Bayes Function Tagger [11] are used. English Language Analyzer [12] is used for English sentence analysis.

3.2. Reordering Rule Generation

In rule generation, reordering rules are automatically extracted from the tagged aligned corpus by using the reordering rule generation algorithm described in Algorithm 3.2. In this work, Myanmar-English reordering rules and English-Myanmar reordering rules are generated. Two basic reordering rules; Part-of-Speech tag based reordering rules and function tag based reordering rules are generated from rule generation algorithm.

Algorithm 3.2: Rule Generation Algorithm

Input: Tagged Aligned Sentences
Output: Reordering Rules
begin
    load tagged aligned copus
    extract each sentence from the copus
    for each tagged aligned sentence
      begin
        extract POS tag reordering rules
        extract function tag reordering rules
      end
end

In English-Myanmar reordering rule generation, there are two findings. The first finding is that there are two or more possible reordering rules for one English sentence because Myanmar is verb final and free word order language. The second finding is that there are also rule ambiguities in Part-of-Speech reordering rules which composed of the Part-of-Speech tag, determiner. Therefore, optimal reordering rules selection from possible reordering rules is solved according to the probability of each reordering rule and Part-of-Speech reordering rule ambiguity is solved by using the lexical information of the Part-of-Speech tag.

3.3. Reordering

In this work, reordering is performed in two hierarchical levels; word level and chunk level. For word level reordering, Part-of-Speech reordering rules obtained from morphological analysis are used. For chunk level reordering, function tag reordering rules obtained from syntactic analysis are applied. The step by step procedure for reordering is performed according to the reordering algorithm, Algorithm 3.3.

Algorithm 3.3: Reordering Algorithm

Input: English sentences or Myanmar sentences

Output: Reordered sentences
begin
accept the input sentence
analyze the input sentence
extract the syntactic structure
extract function tag reordering rule for this syntactic structure
search the possible reordering rules for this structure
select the optimal reordering rule based on the maximum probability
reorder the chunks in the sentence by optimal function tag reordering rule
for each chunk in the sentence
begin
extract possible POS reordering rules
select optimal POS reordering rule based on maximum probability
reorder words in this chunk by optimal POS reordering rule
end
output the reordered sentence
end

In this reordering, optimal reordering rule selection is done by the following equation.

$$P(r_1^n / p_1^n) = \frac{count(r_1^n)}{count(p_1^n)} \quad (1)$$

where, $count(r_1^n)$ is number of reordering suggestion for one reordering pattern and $count(p_1^n)$ is the number of occurrence of this reordering pattern in the corpus.

4. Evaluation of Proposed Reordering System Using 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. In this work, the quality of word order is measured by using permutation distance metrics and lexical reordering metrics. The step by step evaluation procedure is performed according to the Algorithm 4.1.

Algorithm 4.1: Evaluation Algorithm

begin
load bilingual sentences
extract alignment matrix for these sentences
change alignment matrix into permutation matrix using the bi-jective function
$a:\{i \rightarrow j\}$
extract permutation metric for reordered sentence
calculate Hamming distance and Kendall Taue distance
calculate BLEU score and LRscore
calculate overall accuracy for permutation metrics and lexical reordering metric, LRscore
end

In order to evaluate the proposed reordering system according to the evaluation algorithm 4.1, Hamming distance calculation is calculated by using the equation shown in equation 2.

$$d_h(\pi,\sigma) = 1 - \frac{\sum_{i=n}^{n} x_i}{n}, x_i = \begin{cases} 1 & \text{if } \pi(i) = \sigma(i) \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Similarly, Kendall's Tau distance is calculated by the equation 3. In these two equations, n is the length of permutation.

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

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

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

n = the length of the permutation

(3)

For calculating BLEU score, BLEU equation shown in equation 4 is used.

$$BLEU = BP * \exp(\sum_{n=1}^{N} w_n \log p_n) \quad (4)$$

Where, p_n means the modified n-gram precision, BP means the brevity penalty. Moreover, N is defined over which language model is used. In the proposed system, tri-gram language model is used and so N=3 and $w_n = 1/N$ =1/3.

Finally, the lexicalized reordering score (LRscore) is calculated by the following equation 5.

$$LRscore = \alpha R + (1-\alpha)L \quad (5)$$

The lexical reordering metric contains only one parameter, α, which balances the contribution of the reordering metric, R, and the lexical metric, L. In the equation (5), BLEU is used as the lexical metric; L. R is the average permutation distance metric adjusted by the brevity penalty.

The average percentage of Hamming distance and Kendall's Tau distance calculation is describe in Table 1 and the average percentage of LRscore is shown in Table 2.

Table 1. Average Percentage of Permutation Metrics

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

reference		

In the table 2, the LRscore calculated by Hamming distance and uni-gram BLEU score (LR-HB1), the LRscore calculated by Kendall's Tau distance with uni-gram BLEU score (LR-KB1), the LRscore with Hamming distance and tri-gram BLEU score (LR-HB3) and the LRscore with Kendall's Tau distance and tri-gram BLEU score (LR-KB3) are described. As shown in Table2, the LRscores calculated by using tri-gram BLEU score is higher than the LRscores calculated by uni-gram BLEU score.

Table 2. Average Percentage of Permutation Metrics

	LR-HB1	LR-HB3	LR-KB1	LR-KB3
LRscore (%)	68.5	70.1	72.5	74.1

5. Experimental Results

Proposed reordering system is tested over 3000 English sentences and 3000 Myanmar sentences. Experimental results of proposed system over different sentence type are explained in section 4.1.

5.1. Evaluation of Proposed System over Different Sentence Types

In the proposed reordering system, three types of English sentences, simple, compound, and complex sentences, and two types of Myanmar sentences, simple and complex Myanmar sentences are trained and tested. The minimum word length of English sentences is 3 and the maximum word length is 25. The minimum word length of Myanmar sentences is 6 and the maximum word length is 20. When evaluation is carried over these types of sentences, it can be seen that the sentences which have long word length may have more errors than those of short word length. The accuracy of proposed system over the sentences is shown as a bar chart in the Figure 2 and Figure 3 respectively.

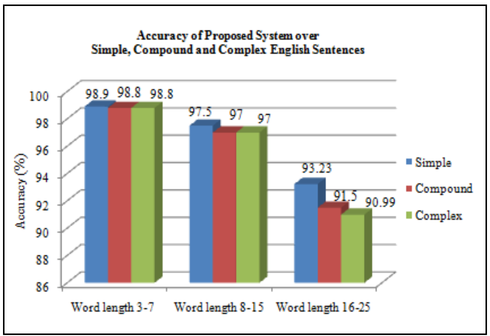

Figure 2. Accuracy of English-Myanmar Reordering

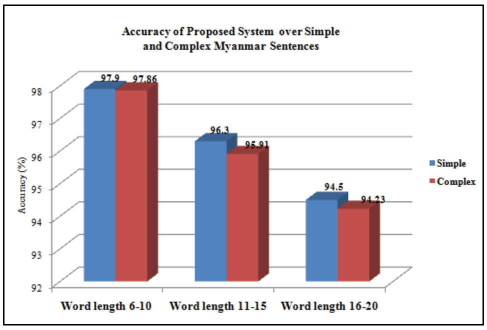

Figure 3. Accuracy of Myanmar-English Reordering

6. Conclusion and Future Work

In the proposed reordering system, there are three main contributions; such as tagged aligned corpus creation, automatic reordering rules generation, and performing word-level and chunk-level reordering based on the reordering rules which have maximum probability. Tagged aligned corpus creation is carried out by analyzing the plain text corpus and this corpus has root words, chunk type, function tag, POS tag, and alignment positions. Therefore, this corpus can be used in another process such as information retrieval other than rule generation and this corpus creation procedure can be used in building the tagged aligned corpus for other languages.

Moreover, in this reordering system, automatic Part-of-Speech and function tag reordering rule generation algorithm is also proposed. This rule generation algorithm uses the tagged aligned corpus as a resource and therefore it can be used for generating reordering rules for every language which have tagged aligned corpus. In addition to this, the proposed reordering system performs reordering in two hierarchical levels; word level and chunk level and thus long-distance reordering can be solved.

Therefore, the working style of proposed reordering system can be applied in other language pairs which need long-distance reordering. In this research work, English sentences which have word length 3 to 25 and Myanmar sentences with word length 7 to 20 are trained and tested. As the future work, the proposed reordering system will be extended for more complex and longer English sentences and Myanmar sentences by adding the reordering patterns of these sentences into the reordering rule base.

References

[1] Y. Al-Onaizan and K. Papineno. "Distortion models for statistical machine translation." In Proceedings of the 21st International Conference on Computational Linguistics and the 4th annual meeting of the ACL, pp. 529–536, Sydney, Australia.

[2] J.V. Graca, K.Ganchev, and Ben Taskar. "Learning Tractable Word Alignment Models with Complex Constraints."

[2] S. Vogel, F.J. Och, C. Tillmann, S. Nießen, H. Sawaf, and H.Ney. "Statistical methods for machine translation." pp.377–393. Springer Verlag: Berlin, Heidelberg, New York, 2000.
[3] C. Tillmann and H. Ney. "Word reordering and DP beam search for statistical machine translation to appear in Computational Linguistics." NATO ASI Series F68, Berlin: Springer Verlag, pp. 227-236, 2000.
[4] L. Shen, A. Sarkar, and F. J. Och. "Discriminative reranking for machine translation.", 2004.
177.
[5]Z. Zheng, Y. H. Yu "Lexical-based Reordering Model for Hierarchical Phrase-based Machine Translation"
[6]B. Chen, M. Cettolo, and M. Federico. "Reordering rules for phrase-based statistical machine translation.", In International Workshop on Spoken Language Translation Evaluation Campaign on Spoken Language Translation, pages 1–15, 2006.
[7] M. Popovic and H. Ney. " POS-based word reorderings for statistical machine translation." Genoa,
Italy,2006.
[8] Y. Zhang, R. Zens, and H. Ney. "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. 2007.
[9]W.P.Pa and N.L.Thein, "Myanmar Word Segmentation using Hybrid Approach." In Proc. 7th International Conference for Computer Application. Yangon, Myanmar, May 5-6, 2009.
[10] P.H.Myint, T.M.Htwe and N.L.Thein, "Bigram Part-of-Speech Tagger for Myanmar Language" International Conference on Information Communication and Management, IPCSIT vol.16(2011) , IACSIT Press (2011), Singapore.
[11] W.W.Thant, T.M.Htwe and N.L.Thein, "Syntactic Analysis of Myanmar Language", Proceedings of International Conference on Computer Applications (ICCA 2011), Yangon, Myanmar, May 5-6, 2011.
[12] M.T. Tun and N.L.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.