A Large-scale Study of Statistical Machine Translation Methods for Myanmar Language

Ye Kyaw Thu†, Andrew Finch†, Win Pa Pa‡, and Eiichiro Sumita†

†Advanced Speech Translation Research and Development Promotion Center,
NICT, Kyoto, Japan

‡Natural Language Processing Lab,
University of Computer Studies, Yangon, Myanmar

{yekyawthu,andrew.finch,eiichiro.sumita}@nict.go.jp
winpapa@ucsy.edu.mm

Abstract. This paper contributes the first large scale evaluation of the quality of automatic translation between Myanmar and twenty other languages, in both directions. The experiments were carried out using three different statistical machine translation approaches: phrase-based, hierarchical phrase-based, and the operation sequence model (OSM). In addition three different segmentation schemes for Myanmar were studied, these were syllable segmentation, maximum matching word segmentation with a dictionary and supervised word segmentation. The results show that the highest quality machine translation was attained with supervised word segmentation in all of the experiments. Furthermore, for almost all language pairs the HPBSMT approach gave the highest translation quality when measured in terms of both the BLEU and RIBES scores.

Keywords: Machine translation, Myanmar, Operation Sequence Model, Hierarchical Phrase-based, Word Segmentation

1 Introduction

The main contribution of this paper, is the first large-scale study of Myanmar statistical machine translation. Myanmar machine translation is still in its early stages and researchers are faced with many difficulties arising from the lack of resources, in particular parallel corpora are scarce. Furthermore, the techniques for performing requisite pre-processing for Myanmar language, such as word segmentation are also currently in the process of being developed. Existing research on Myanmar translation has been either rule-based (which avoids the issue of scarce resources) [21] or more recently phrase-based [22] techniques have been tried. Advanced approaches such as the hierarchical phrase-based approach and the operation sequence model have so far received no attention in the literature, but these approaches offer the promise of better being able to meet the challenges posed by word re-ordering. The Myanmar language although similar in terms of word order to languages such as Japanese and Korean is dissimilar in word order to many important of languages, in particular the Indo-European family of

languages which accounts for around 45% of the world's speakers. In this paper we study for the first time the application of these advanced approaches to the translation of the Myanmar language.

The structure of the paper is as follows. In the next section we briefly introduce the Myanmar language, outline the approaches taken so far to Myanmar word segmentation, and describe the three approaches we have chosen to examine in this study. These are a simple approach that divides Myanmar into its component syllables, a maximum matching method with a dictionary, and a more sophisticated supervised word segmentation approach. Then, we describe the methodology used in the machine translation experiments, present the results of these experiments, and finally conclude.

1.1 Myanmar Language

In Myanmar text, words composed of single or multiple syllables are usually not separated by white space. Although spaces are used for separating phrases for easier reading, it is not strictly necessary, and these spaces are rarely used in short sentences. There are no clear rules for using spaces in Myanmar language, and thus spaces may (or may not) be inserted between words, phrases, and even between root words and their affixes.

2 Related Work

The problems of Myanmar word segmentation have been analyzed and different approaches have been developed to achieve different goals.

[17] proposed a hybrid approach that works by longest matching on syllable-segmented sentences. Their probabilistic model used a lexicon of 20,000 words from a Myanmar grammar [14] and achieved 0.755 precision.

[20] proposed a word segmentation approach that involved rule-based syllable segmentation and dictionary-based statistical syllable merging using a dictionary of about 30,000 words provided by the Myanmar NLP team of the Myanmar Computer Federation. Their approach achieved 100% syllable accuracy and 98.94% precision, 99.05% recall and 98.99% F-score on their word segmentation task.

[6] proposed a 2-step longest matching approach. The first step, was syllable segmentation, in the second step left-to-right syllable longest matching forward segmentation was performed. A 2 million sentence monolingual Myanmar corpus and an 80K sentence English-Myanmar parallel corpus, and lists of stop words, syllables and words were used in the decision process for annotating word boundaries. This approach employed a similar longest matching strategy to [17], and as a consequence also suffers the same problems relating to ambiguity.

[22] studied word segmentation in the context of statistical machine translation using 7 different schemes, including a proposed unsupervised segmentation approach which did not exceed the performance of the simpler maximum matching approach. They hypothesized that the cause was a lack of data.

In this work we focused on a supervised approach which we expected to perform well training on a small amount of human segmented data.

To date, there have been very few studies on the automatic translation of Myanmar language machine. In [21] a method for word to phrase re-ordering for Myanmar-English translation based on English grammar rules was proposed. [24] studied Myanmar word disambiguation for Myanmar-English SMT. In [26], a Myanmar phrase translation model with morphological analysis for Statistical Myanmar to English translation. All previous research has been based on very small parallel corpora (the largest being 13,042 sentence pairs). To the best of our knowledge, there exists no study on the same scale as the experiments reported here.

3 Segmentation

3.1 Syllable Segmentation

Syllable breaking is a necessary step for Myanmar word segmentation, since most Myanmar words are composed of multiple syllables and most of the syllables are composed of more than one character. We used the algorithm of [22] for syllable breaking. There are three general rules to break Myanmar syllables from Unicode input text where a consonant is followed by dependent vowels and other symbols.

The first rule puts a word break in front of consonants, independent vowels, numbers and symbol characters. The second rule removes any word breaks that are in front of subscript consonants, Kinzi characters, and consonant + Asat characters. For the break points in special cases such as syllable combinations of loan words (e.g. အဲလက်စ် , that is the transliteration of "Alex"), Pali words (e.g. တက္က သိုလ်, university in English), we used orthographic segmentation. In experiments for these rules with a 27,747 word dictionary the approach was able to achieve 100% precision and recall [22].

3.2 Maximum Matching

Maximum matching is one of the most popular structural segmentation algorithms and it is often used as a baseline method in word segmentation [25]. This method segments using segments chosen from a dictionary. The method strives to segment using the longest possible segments. It is a greedy algorithm and is therefore sub-optimal. The segmentation process may start from either end of the sequences. In this paper, we used left-to-right maximum matching using modified Myanmar Language Commission (MLC) dictionary [4] that contained 24,562 words. Here, modified means we removed one syllable words from the original MLC dictionary for getting bigger unit than syllables.

3.3 Conditional Random Fields

Linear-chain conditional random fields (CRFs) [13] are models that consider dependencies among the predicted segmentation labels that are inherent in the state transitions of finite state sequence models and can incorporate domain knowledge effectively into the segmentation process. Unlike heuristic methods, they are principled probabilistic finite state models on which exact inference

over sequences can be efficiently performed. The model computes the following probability of a label sequence $\mathbf{Y} = \{y_1, ..., y_T\}$ of a particular character string $\mathbf{W} = \{w_1, ..., w_T\}$.

$$P_{\boldsymbol{\lambda}}(\mathbf{Y}|\mathbf{W}) = \frac{1}{Z(\mathbf{W})} exp(\sum_{t=1}^{T} \sum_{k=1}^{|\boldsymbol{\lambda}|} \lambda_k f_k(y_{t-1}, \mathbf{W}, t)) \quad (1)$$

where $Z(\mathbf{W})$ is a normalization term, f_k is a feature function, and $\boldsymbol{\lambda}$ is a feature weight vector.

We used the CRF++ toolkit [12] to build the CRF models. The feature set used in the models (up to character/syllable tri-grams) was as follows (where t is the index of the character/syllable being labeled):

- unigrams:
  $\{w_{t-2}, w_{t-1}, w_t, w_{t+1}, w_{t+2}\}$
- bigrams:
  $\{(w_{t-1}, w_t), (w_t, w_{t+1})\}$
- trigrams:
  $\{(w_{t-2}, w_{t-1}, w_t), (w_{t-1}, w_t, w_{t+1}),$
  $(w_t, w_{t+1}, w_{t+2})\}$

These n-grams were combined with label unigrams and bigrams to produce the feature set for the model.

4 Experimental Methodology

4.1 Corpus Statistics

We used twenty languages from the multilingual Basic Travel Expressions Corpus (BTEC), which is a collection of travel-related expressions [9]. The languages were Arabic (ar), Chinese (zh), English (en), German (de), Hindi (hi), Indonesian (id), Italian (it), Japanese (ja), Korean (ko), Malaysian (ms), Mongolian (mn), Myanmar (my), Nepali (ne), Portuguese (br), Russian (ru), Sinhala (si), Spanish (es), Tagalog (tl), Thai (th), Turkish (tl) and Vietnamese (vi). This resulted on 40 language pairs being used in the experiments. 457,249 sentences were used for training, 5,000 sentences for development and 3,000 sentences for evaluation.

In all experiments, the Myanmar language was segmented using rule based syllable segmentation, maximum matching, and the CRF word segmentation methods described in Sections 3.1, 3.2 and 3.3 respectively.

4.2 Phrase-based Statistical Machine Translation (PBSMT)

We used the phrase based SMT system provided by the Moses toolkit [10] for training the phrase-based machine statistical translation system. The Myanmar was aligned with the word segmented target languages using GIZA++ [15]. The alignment was symmetrized by grow-diag-final-and heuristic [11]. The lexicalized reordering model was trained with the msd-bidirectional-fe option [23]. We use SRILM for training the 5-gram language model with interpolated modified Kneser-Ney discounting [19]. Minimum error rate training (MERT) [16] was used to tune the decoder parameters and the decoding was done using the Moses decoder (version 2.1) [10].

4.3 Hierarchical Phrase-based Machine Translation (HPBSMT)

The hierarchical phrase-based SMT approach [3] is a model based on synchronous context-free grammar. The models are able to be learned from a corpus of unannotated parallel text. The advantage this technique offers over the phrase-based approach is that the hierarchical structure is able to represent the word re-ordering process. The re-ordering is represented explicitly rather than encoded into a lexicalized re-ordering model (commonly used in purely phrase-based approaches). This makes the approach particularly applicable to languages pairs that require long-distance re-ordering during the translation process [2]. For the experiments in this paper we used the implementation of hierarchical model provided by the Moses machine translation toolkit (both the hierarchical decoder and training procedure provided by the experiment management system), using the default settings.

4.4 Operation Sequence Model (OSM)

The operation sequence model is a model for SMT that combines the benefits of two state-of-the-art SMT frameworks, namely n-gram-based SMT and phrase-based SMT [5]. It is a generative model that performs the translation process as a linear sequence of operations that jointly generate the source and target sentences. The operation types are (i) generation of a sequence of source and/or target words (ii) insertion of gaps as explicit target positions for reordering operations, and (iii) forward and backward jump operations which perform the actual reordering. The probability of a sequence of operations is given by an n-gram model. The OSM integrates translation and reordering into a single model which provides a natural reordering mechanism that is able to correctly re-order words across long distances. We used Moses [10] for training the OSM, with n-gram model order 5. Other settings such as those used to build the language model and lexicalized reordering model were the same as the default PBSMT system (refer to Section 4.2 for details).

4.5 Evaluation Criteria

We used two automatic criteria for the evaluation of the machine translation output. One was the de facto standard automatic evaluation metric Bilingual Evaluation Understudy (BLEU) [18] and the other was the Rank-based Intuitive Bilingual Evaluation Measure (RIBES) [7]. The BLEU score measures the precision of n-grams (over all $n \leq 4$ in our case) with respect to a reference translation with a penalty for short translations [18]. Intuitively, the BLEU score measures the adequacy of the translations and large BLEU scores are better. RIBES is an automatic evaluation metric based on rank correlation coefficients modified with precision and special care is paid to word order of the translation results. The RIBES score is suitable for distant language pairs such as Myanmar and English, Myanmar and Thai, Myanmar and Vietnamese [7]. Large RIBES scores are better. For those language pairs in which the target language was Myanmar, the translations were decomposed into their constituent syllables in order to ensure the results were cross-comparable.

Src-Trg	Syllable			Word (Max-Match)			Word (CRF)		
	PBSMT	HPBSMT	OSM	PBSMT	HPBSMT	OSM	PBSMT	HPBSMT	OSM
my-ar	20.60 (0.63)	27.90 (0.68)	20.10 (0.61)	26.87 (0.67)	32.55 (0.71)	26.41 (0.67)	32.00 (0.70)	**34.16** (**0.72**)	32.09 (0.70)
my-br	29.21 (0.73)	37.88 (0.79)	29.17 (0.74)	35.57 (0.78)	40.56 (0.81)	35.80 (0.78)	39.12 (0.80)	**41.52** (**0.82**)	39.45 (0.80)
my-de	26.82 (0.74)	31.87 (0.78)	27.08 (0.74)	32.29 (0.77)	**35.90** (**0.79**)	32.78 (0.77)	34.98 (0.78)	35.80 (0.79)	35.14 (0.78)
my-en	33.14 (0.76)	42.76 (0.83)	33.31 (0.77)	40.28 (0.81)	45.86 (0.85)	39.83 (0.81)	43.82 (0.82)	**46.97** (**0.85**)	44.46 (0.83)
my-es	28.54 (0.72)	39.01 (0.79)	28.79 (0.73)	35.72 (0.78)	**42.69** (**0.82**)	35.35 (0.78)	39.49 (0.79)	42.02 (0.81)	40.08 (0.79)
my-hi	29.44 (0.70)	30.86 (0.71)	30.29 (0.70)	33.05 (0.71)	**33.87** (**0.73**)	33.45 (0.73)	33.31 (0.72)	33.71 (0.72)	33.74 (0.72)
my-id	29.63 (0.76)	39.25 (0.82)	29.86 (0.76)	36.28 (0.80)	42.04 (**0.84**)	36.55 (0.80)	41.38 (0.81)	**43.96** (0.83)	41.62 (0.82)
my-it	26.77 (0.70)	33.87 (0.74)	26.94 (0.70)	32.96 (0.75)	**37.57** (**0.78**)	33.39 (0.75)	35.69 (0.76)	37.41 (0.77)	35.96 (0.76)
my-ja	34.28 (0.79)	34.46 (0.79)	34.28 (0.79)	35.42 (0.79)	**35.77** (0.79)	35.50 (0.80)	35.36 (**0.80**)	35.36 (0.79)	35.57 (0.79)
my-ko	29.95 (0.74)	30.33 (0.74)	30.32 (0.74)	31.72 (0.75)	32.27 (0.76)	32.36 (0.75)	32.15 (**0.76**)	32.44 (0.76)	**33.03** (0.76)
my-mn	30.43 (0.74)	30.86 (0.74)	30.71 (0.75)	33.88 (0.77)	33.52 (0.76)	34.29 (0.77)	35.52 (0.77)	35.93 (0.77)	**36.83** (**0.77**)
my-ms	26.85 (0.73)	34.99 (0.80)	26.78 (0.74)	33.52 (0.78)	38.15 (**0.81**)	33.34 (0.78)	36.78 (0.79)	**38.58** (0.81)	36.57 (0.79)
my-ne	28.32 (0.73)	29.09 (0.73)	29.03 (0.74)	31.75 (0.75)	32.00 (0.75)	31.95 (0.75)	33.13 (0.76)	**33.69** (0.76)	33.52 (**0.76**)
my-ru	24.47 (0.67)	31.30 (0.71)	24.90 (0.68)	29.88 (0.72)	34.29 (**0.75**)	30.37 (0.73)	33.22 (0.73)	**34.68** (0.74)	33.51 (0.73)
my-si	28.17 (0.70)	28.81 (0.70)	28.72 (0.71)	31.63 (0.72)	32.29 (0.73)	31.81 (0.72)	33.20 (0.73)	**34.06** (0.74)	33.53 (**0.74**)
my-th	26.44 (0.62)	33.05 (0.70)	26.12 (0.61)	31.06 (0.67)	35.64 (**0.73**)	31.36 (0.67)	34.10 (0.70)	35.39 (0.73)	**33.65** (0.70)
my-tl	23.87 (0.70)	31.66 (0.76)	23.77 (0.70)	29.37 (0.74)	34.10 (0.77)	29.61 (0.74)	33.37 (0.76)	**35.63** (**0.78**)	33.21 (0.76)
my-tr	29.69 (0.72)	30.94 (0.72)	30.69 (0.72)	33.19 (0.75)	34.80 (0.76)	34.01 (0.75)	36.34 (**0.77**)	**37.19** (0.77)	36.63 (0.77)
my-vi	32.48 (0.77)	39.59 (0.82)	33.01 (0.77)	37.93 (0.81)	**42.43** (**0.84**)	37.98 (0.81)	40.35 (0.82)	41.74 (0.84)	40.13 (0.82)
my-zh	23.85 (0.70)	25.35 (0.71)	23.79 (0.69)	25.80 (0.71)	27.10 (0.72)	26.17 (0.72)	26.54 (0.72)	**27.65** (0.72)	27.06 (**0.73**)

Table 1: BLEU and RIBES scores for translating from Myanmar.

Src-Trg	Syllable			Word (Max-Match)			Word (CRF)		
	PBSMT	HPBSMT	OSM	PBSMT	HPBSMT	OSM	PBSMT	HPBSMT	OSM
ar-my	36.89 (0.82)	34.83 (0.81)	37.09 (0.82)	39.75 (0.83)	39.27 (0.83)	39.63 (0.83)	40.95 (0.83)	**41.63** (**0.84**)	41.06 (0.84)
br-my	38.11 (0.83)	38.66 (0.84)	38.40 (0.83)	40.26 (0.84)	41.62 (**0.85**)	40.38 (0.83)	41.85 (0.84)	**44.04** (0.84)	42.30 (0.84)
de-my	35.83 (0.82)	37.21 (0.83)	37.24 (0.83)	39.97 (0.84)	40.87 (0.85)	40.64 (0.84)	41.01 (0.84)	**42.16** (**0.85**)	41.10 (0.84)
en-my	39.46 (0.83)	41.22 (0.85)	40.95 (0.84)	42.96 (0.78)	44.15 (0.87)	42.40 (0.85)	44.87 (0.85)	**46.28** (**0.87**)	45.18 (0.86)
es-my	37.60 (0.82)	38.23 (0.84)	38.00 (0.82)	40.60 (0.83)	42.50 (0.85)	40.56 (0.83)	41.88 (0.84)	**44.03** (**0.86**)	41.69 (0.84)
hi-my	39.52 (0.84)	38.84 (0.84)	40.60 (0.85)	42.78 (0.86)	43.02 (0.85)	43.75 (0.86)	43.52 (0.86)	43.56 (0.85)	**43.80** (**0.86**)
id-my	38.39 (0.83)	38.05 (0.84)	39.60 (0.84)	41.14 (0.84)	42.21 (0.85)	42.00 (0.85)	43.59 (0.85)	**45.07** (**0.86**)	43.82 (0.85)
it-my	37.21 (0.82)	37.79 (0.84)	37.73 (0.83)	39.75 (0.83)	41.15 (**0.85**)	40.49 (0.83)	41.07 (0.83)	**41.57** (0.85)	41.50 (0.84)
ja-my	33.00 (0.82)	33.41 (0.82)	33.56 (0.82)	35.24 (0.82)	35.37 (0.82)	35.61 (0.82)	35.88 (0.82)	**36.57** (**0.82**)	35.91 (0.82)
ko-my	33.72 (0.83)	33.66 (0.83)	34.68 (0.83)	36.42 (0.83)	36.60 (0.83)	36.22 (0.83)	37.43 (0.83)	**37.94** (**0.84**)	37.14 (0.83)
mn-my	35.80 (0.83)	35.53 (0.83)	36.97 (0.84)	39.22 (0.85)	39.18 (0.84)	40.31 (0.85)	41.83 (**0.85**)	41.46 (0.85)	**42.11** (0.85)
ms-my	37.47 (0.83)	37.88 (0.84)	38.20 (0.83)	40.46 (0.84)	41.89 (0.85)	41.08 (0.84)	42.69 (0.85)	**43.86** (**0.86**)	42.55 (0.85)
ne-my	38.45 (0.84)	38.13 (0.83)	40.46 (0.85)	41.94 (0.85)	41.59 (0.85)	42.69 (0.85)	**43.31** (0.85)	43.04 (0.85)	43.09 (**0.85**)
ru-my	35.93 (0.82)	35.35 (0.82)	36.84 (0.82)	38.22 (0.82)	38.68 (0.83)	38.80 (0.82)	39.17 (0.82)	**39.86** (**0.84**)	39.42 (0.83)
si-my	39.94 (0.84)	39.43 (0.84)	40.76 (0.71)	42.04 (0.80)	42.11 (0.84)	43.60 (0.80)	43.10 (0.85)	**43.90** (**0.85**)	43.59 (0.85)
th-my	34.26 (0.79)	35.81 (0.82)	33.80 (0.79)	36.44 (0.80)	38.71 (0.84)	36.51 (0.80)	38.00 (0.81)	**40.98** (**0.84**)	37.26 (0.81)
tl-my	35.32 (0.80)	36.64 (0.83)	35.84 (0.81)	38.52 (0.82)	40.16 (0.85)	39.13 (0.82)	39.62 (0.82)	**41.47** (**0.85**)	39.46 (0.83)
tr-my	37.42 (0.84)	36.94 (0.85)	38.46 (0.85)	40.79 (0.85)	40.63 (0.86)	41.54 (0.86)	43.01 (0.86)	43.27 (0.86)	**43.79** (**0.87**)
vi-my	36.44 (0.82)	38.19 (0.84)	36.66 (0.82)	38.63 (0.83)	41.64 (0.85)	39.18 (0.83)	39.41 (0.83)	**42.42** (**0.86**)	39.06 (0.83)
zh-my	30.21 (0.79)	29.58 (0.79)	30.28 (0.79)	31.80 (0.79)	31.87 (0.79)	32.12 (0.79)	32.02 (0.80)	**32.53** (**0.80**)	31.88 (0.79)

Table 2: BLEU and RIBES scores for translating to Myanmar.

5 Results

The BLEU and RIBES score results for machine translation experiments with PBSMT, HPBSMT and OSM are shown in Tables 1 (translating from Myanmar) and 2 (translating into Myanmar). Bold numbers indicate the highest scores of the three different approaches.

5.1 Discussion

Looking at the results in Tables 1 and 2, it is clear that the supervised CRF-based word segmentation scheme was by far the most effective. For translation from Myanmar there were a few cases (30%) where the maximum matching segmentation method gave better results in terms of BLEU score, but for translation into Myanmar the CRF approach dominated, outperforming the other word segmentation approaches for all language pairs. For clarity, we will therefore limit our discussion from here on to only those results from the best performing approach: the CRF segmenter.

The highest absolute BLEU scores were achieved on the Myanmar-English and English-Myanmar tasks. We believe that the reason for this is not because it is easy to translate between these two languages, but that most of the Myanmar part of the corpus was created by translating from the English sentences. Another factor that may have contributed here is that the English data (along with the Japanese) has been subjected to the the most checking and revision, and is among the cleanest languages in the corpus.

Our original motivation for studying the application the OSM and HPBSMT techniques would be advantageous for translating between Myanmar and many other languages. From the results in Table 1 it can be seen that the PBSMT approach was not the most effective approach in terms of BLEU for any language pair in our experimental set. The HPBSMT approach was the most effective, giving rise to the highest BLEU score for 75% of the experiments. The character of the results was similar in our experiments translating to Myanmar. Here the PBSMT approach achieved the highest BLEU score for only one language pair. The HPBSMT approach again was the most effective for 80% of the experiments, and the OSM approach give the best score in 15% of the experiments.

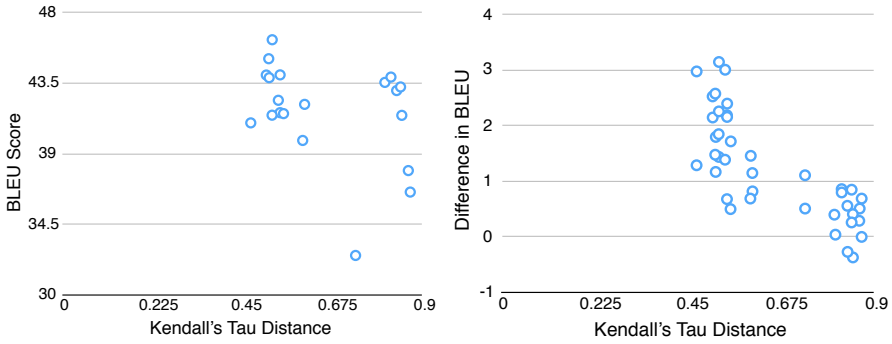
Our hypothesis was the HPBSMT and OSM approaches would give the most benefit when the languages had different word orders, and this indeed seems to have been the case by visual inspection of the results. Looking at the results translating between Japanese and Korean and Myanmar, it can be seen that the gain in BLEU of the HPBSMT system over the baseline PBSMT system is only a fraction of a BLEU point (average 0.37 BLEU), whereas the gains in BLEU for translation between Myanmar and the European languages (English, Spanish, Italian, German) is much larger (average 1.68 BLEU). This motivates a deeper study of this phenomenon that we report in the next section.

5.2 Analysis

We analyzed the results using Kendall's tau distance in order to guage the effect of re-ordering on the translation process.

We calculated the Pearson product-moment correlation coefficient (PMCC) between the BLEU score and the Kendall's tau distance [8] to assess the strength of the linear relationship between the amount of reordering required during the translation process and the translation quality.

Kendall's tau distance is based on the number of transpositions of adjacent symbols necessary to transform one permutation into another [8, 1], and is one method to gauge the amount of re-ordering that would be required during the translation process between two languages. In this paper we use the version defined in [1] in which maximally close permutations have a distance of 1 and maximally distant permutations have a distance of 0.



(a) Plot of the Kendall's tau distance against BLEU.

(b) Plot of the Kendall's tau distance against BLEU differences.

Fig. 1: The effect of re-ordering on translation performance

Figure 1a shows a scatter plot of all of the HPBSMT experiments with CRF segmented Myanmar as the target language, plotting BLEU score against Kendall's tau distance. The data shows a weak negative correlation (coefficient: -0.36), indicating that the degree of re-ordering has some influence on the translation quality. This relationship here however is very surprising, since the languages that are close in terms of word order to Myanmar appear to be more difficult to translate, counter to intuition. As mentioned earlier, we believe the explanation for this lies in the fact that the Myanmar was mainly translated from the English during the creation of the corpus, and this given some advantage in the translation of English (and other closely related European languages) in the experiments.

More importantly, Figure 1b shows a scatter plot of all of the HPBSMT experiments, plotting BLEU score differences against Kendall's tau distance. The data shows a strong negative correlation (coefficient: -0.78), indicating that the degree of improvement gained by using the HPBSMT approach is strongly related to the degree of re-ordering required during translation. This result supports the main hypothesis motivating the work reported in this paper.

6 Conclusion

This paper has presented the first large-scale study of the translation of the Myanmar language. A total of 40 language pairs were used in the study that included languages both similar and fundamentally different from Myamnar. We proposed a method of supervised CRF-based word segmentation which was shown was capable of improving end-to-end machine translation performance relative to a baseline segmentation based on maximum matching, even when trained on only a small data set. In addition, we evaluated the performance of three machine translation techniques on all of the language pairs, these were: phrase-based SMT, hierarchical phrase-based SMT and the operation sequence model. Our results show that the hierarchical phrase-based method was the most effective in general for almost all language pairs achieving gains of over 3 BLEU points for some language pairs. More specifically, our original motivation for this study was that using approaches such as HPBSMT that better handle the re-ordering process would be particularly suitable for Myanmar translation since Myanmar has a vastly different word order to the languages in the large Indo-European family. Our results show that the absolute gains increased with the amount of word re-ordering required during translation.

References

1. Birch, A.: Reordering Metrics for Statistical Machine Translation. Ph.D. thesis, University of Edinburgh (2011)
2. Braune, F., Gojun, A., Fraser, A.: Long-distance reordering during search for hierarchical phrase-based smt. In: EAMT 2012: Proceedings of the 16th Annual Conference of the European Association for Machine Translation, Trento, Italy. pp. 177–184. Citeseer (2012)
3. Chiang, D.: Hierarchical phrase-based translation. Computational Linguistics 33(2), 201–228 (2007)
4. Commission, M.L.: Myanmar English Dictionary. Myanmar Language Commission (2012)
5. Durrani, N., Schmid, H., Fraser, A., Koehn, P., Schütze, H.: The Operation Sequence Model – Combining N-Gram-based and Phrase-based Statistical Machine Translation. Computational Linguistics 41(2), 157–186 (2015)
6. Htay, H.H., Murthy, K.N.: Myanmar word segmentation using syllable level longest matching. In: IJCNLP. pp. 41–48 (2008)
7. Isozaki, H., Hirao, T., Duh, K., Sudoh, K., Tsukada, H.: Automatic evaluation of translation quality for distant language pairs. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. pp. 944–952. EMNLP '10, Association for Computational Linguistics, Stroudsburg, PA, USA (2010), http://dl.acm.org/citation.cfm?id=1870658.1870750
8. Kendall, M.G.: A new measure of rank correlation. Biometrika 30(1/2), 81–93 (1938), http://www.jstor.org/stable/2332226
9. Kikui, G., Sumita, E., Takezawa, T., Yamamoto, S.: Creating corpora for speech-to-speech translation. In: Proceedings of EUROSPEECH-03. pp. 381–384 (2003)
10. Koehn, P., Haddow, B.: Edinburgh's Submission to all Tracks of the WMT2009 Shared Task with Reordering and Speed Improvements to Moses. In: Proceedings of the Fourth Workshop on Statistical Machine Translation. pp. 160–164 (2009)

11. Koehn, P., Och, F.J., , Marcu, D.: Statistical phrase-based translation. In: In Proceedings of the Human Language Technology Conference. Edmonton, Canada (2003)
12. Kudo, T.: Crf++ an open source toolkit for crf, http://crfpp.sourceforge.net/
13. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning. pp. 282–289. ICML '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2001), http://dl.acm.org/citation.cfm?id=645530.655813
14. Myanmar Language Commission, .: Myanmar Thadda (in Myanmar language). Ministry of Education, Yangon, Myanmar, combination of 3 volumes, 1st edn. (2005)
15. Och, F.J., Ney, H.: Improved statistical alignment models. In: ACL00. pp. 440–447. Hong Kong, China (2000)
16. Och, F.J.: Minimum error rate training for statistical machine translation. In: Proceedings of the 41st Meeting of the Association for Computational Linguistics (ACL 2003). Sapporo, Japan (2003)
17. Pa, W.P., Thein, N.L.: Myanmar word segmentation using hybrid approach. Proceeding of the 6th International Conference on Computer Applications pp. 166–170 (2008)
18. Papineni, K., Roukos, S., Ward, T., Zhu, W.: Bleu: a Method for Automatic Evaluation of Machine Translation. IBM Research Report rc22176 (w0109022), Thomas J. Watson Research Center (2001)
19. Stolcke, A.: SRILM - An Extensible Language Modeling Toolkit. In: Proceedings of the International Conference on Spoken Language Processing. vol. 2, pp. 901–904. Denver (2002)
20. Thet, T.T., Na, J.C., Ko, W.K.: Word segmentation for the myanmar language. J. Information Science 34(5), 688–704 (2008)
21. Thida Win, A.: Words to phrase reordering machine translation system in myanmar-english using english grammar rules. In: Computer Research and Development (ICCRD), 2011 3rd International Conference on. vol. 3, pp. 50–53 (March 2011)
22. Thu, Y.K., Finch, A., Sagisaka, Y., Sumita, E.: A study of myanmar word segmentation schemes for statistical machine translation. Proceeding of the 11th International Conference on Computer Applications pp. 167–179 (2013)
23. Tillmann, C.: A unigram orientation model for statistical machine translation. In: Proceedings of HLT-NAACL 2004: Short Papers. pp. 101–104. HLT-NAACL-Short '04, Association for Computational Linguistics, Stroudsburg, PA, USA (2004), http://dl.acm.org/citation.cfm?id=1613984.1614010
24. Wai, T.T., Htwe, T.M., Thein, N.L.: Article: Automatic reordering rule generation and application of reordering rules in stochastic reordering model for english-myanmar machine translation. International Journal of Computer Applications 27(8), 19–25 (August 2011), full text available
25. Yuan Liu, Q.T., Shen, K.X.: The Word Segmentation Methods for Chinese Information Processing (in Chinese). Quing Hua University Press and Guang Xi Science and Technology Press (1994)
26. Zin, T.T., Soe, K.M., Thein, N.L.: Translation model of Myanmar phrases for statistical machine translation., pp. 235–242. Berlin: Springer (2012)