Factored Machine Translation for Myanmar to English, Japanese and Vice Versa

Ye Kyaw Thu†, Andrew Finch†, Akihiro Tamura†, EiichiroSumita† and Yoshinori Sagisaka‡

†*National Institute of Information and Communications Technology*
‡GITI, Speech Science Research Lab., Waseda University
{yekyawthu, andrew.finch, akihiro.tamura, eiichiro.sumita}@nict.go.jp
ysagisaka@gmail.com

Abstract

Factored machine translation models extend traditionalPhrase Based Statistical Machine Translation (PB-SMT) by taking into account notonly the surface form of the words, but also linguistic knowledge such as the dictionary form (lemma), part-of-speech (POS) and morphological tags. In this paper, we used POS tagsas a factor and built factored machine translation models with various translation configurations for Myanmar to English, Japanese translation and vice versa.The experimental results show an improvement in translation quality for Myanmar to English and Myanmar to Japanese language pairs.

1. Introduction

Although factored models have been successfully used in several rich resource language pairs to improve translation quality, to the best of our knowledge,there is no prior research targeting the application of factored models to the machine translation of low resource languages. In this paper, we attempt to increase machine translation performance for a low resource language, Myanmar, by applying a factored machine translation approachusing POS tags as a factor.

2. Overview

As in standard PB-SMT, the main source of data for training a factored model is a parallel corpus. However, the corpus can be factored; each word or token can be tagged with arbitrary linguistic information such as lemma, POS, morphology and word class [1]. For this reason, the phrase mappings can be extracted for each factor and broken up into steps such as translation step (on the phrasal level), generation step (for each word). Each step is modeled by one or more feature functions that are combined by means of a log-linear model with a weight for each feature. Furthermore, order of mapping steps is chosen to optimize search.

In this paper our focus is on low-resource languages. We expect factored models to be effective here, since factors can mitigate issues arising from data sparseness. On the other hand, thereis also lack of data for the factors themselves. Therefore this study aims to provide some insight into how these two factors interact, in other words, can factored models be effective for low resource languages, in spite of incomplete/inaccurate labeling of the factors themselves.

3. Related Work

Koehn and Hoangcarried out factored machine translation experiments on English-German, English-Spanish, English-Czech and English-Chinese language pairs by using linguistic information and automatically generated word classes [1]. They used 20,000 to 750,000 sentences of legal and journalistic domain training data and reported gain up to 2% in BLEU scores over a baseline PB-SMT system without factors. Another factored machine translation experiment done on English-Czech language pair (with acorpus twice as large as the one used in [1]) was reported byBojar, O.[2]. Their results also show that factored machine translation outperformed a traditional PB-SMT approach.Moreover,similar gains could be achievedas the corpus size was increased [3]. The factored machine translation between Brazilian Portuguese and English was studied in [4]. They reported that factored translation models built with POS and morphological information achieved

better results than a system without factored models, but the same gain in performance was not achieve when flat syntactic tags derived from parse tree information were used.Experiments carried out with factored models in both directions for a Russian-English translation task[5] found gains in both directions.

Motivated by the successes in the prior research above, we set out to investigate the effectiveness of factored machine translation for low resource languages where the parallel training data, and data to annotate the factors are both in short supply.

4. Methodology

In this section, we will present pre-processing steps of POS tagging and language model building for factored translation.

For parallel data, we used English, Japanese and Myanmar language data from the multilingual Basic Travel Expressions Corpus (BTEC), which is a collection of travel-related expressions [6]. In this paper, we used 131,698 sentences for training, 20,000 sentences for development and 4,341 sentences for testing.

4.1. POS Tagging

The original English BTEC datawere POS taggedusing the TREE-TAGGER [7]. We used the MeCab Part-of-Speech and Morphological Analyzer for Japanese POS tagging [8].

For Myanmar, both word segmentation and POS tagging are necessary because the language is syllabic and therefore naturally not segmented into words. We used 2,713 Myanmar sentences (24,129 words)tagged with UCSY (University of Computer Studies, Yangon) POS tags data for word segmentation and POS tagging of the Myanmar part of the BTEC corpus [9].A Maximum Matching Word segmentation method was used with unique 2,478 wordsextracted from the UCSY POS tagged data [10], [11].We then used the KyTea text analyzing toolkit for building a model for POS tagging with UCSY POS tags [12].
Examples of POS tagged data are given below, in the format “word|POS-tag”:

English:
How|NP much|RB is|VBZ it|PP ?|SENT

Japanese:
お|接頭詞いくら|名詞です|助動詞か|助詞。|記号

Myanmar:
□□□|PRN.Question □□□□□|Part.Support
□□|Part.Support □□|SF.Interrogative □|UNK

There are many “UNK” (Unknown) tags contained in the POS tagged Myanmar data. 30.57% of the tokens in the training data were labeled with UNK; 30.63% inthe development data, and 31.36% in the test data.

As mentioned earlier, these POS tags are not only incompletely annotated but also can be expected to be inaccurate due to the limited amount of training data available to train the models used to annotate them. This is to be expected in a low-resource language, and we proceed with the experiments in spite of this obstacle.

4.2. Language Models

Modified Kneser-Ney discounted language models (LM) were built from monolingual corpora using IRSTLM toolkit version 5.80.01 [13].Along with the regular 3-gram LM based on word forms, we also used a second 3-gram LM that was trained on POS tags.

5. Experiments and Results

All parallel corpora (Myanmar-Japanese, Japanese-Myanmar, Myanmar-English, English-Myanmar) were aligned with thegrow-diag-final-and heuristic from GIZA++ alignments [14].Both phrase-based translation models and factored translation models were built with MOSES [15].All of the log-linear model weights were optimized on development data using the MERT algorithm [16].The decoding for PBSMT system (for both baselines) and the factored SMT (for factored translations) was done using MOSES version 0.91 [15].

In our experiments, the translation performancewas measured by using a case-sensitive BLEU metric [17]. The significance testing was performed by paired bootstrap re-sampling and computed with the mteval-v13a.pl script provided by NIST [18], [19].

5.1. Baseline Systems

We used two baselines;one was a typical PB-SMT and the otherwas PB-SMT with POS tags annotated

to both source and target words.For both baseline systems, we built a translation model using 131,698 parallel sentence pairs. Trigram language models were builtusing training data of English (987,221 words), Japanese (1,128,425 words) and Myanmar (1,368,710 words).

In the first baseline system (we will call PB-SMT), each word is represented by its surface form; the translation architecture from source to target language is as in Figure 1.

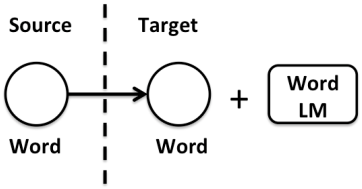

Figure 1. Translation in PB-SMT

The second baseline (PB-SMT-POS) was identical to the first, but used POS tagged words (words/POS) instead of the surface form as tokens during the translation process. The language model wasalso built using POS tagged words. The translation architecture this baseline is shown in Figure2. This baseline incorporates similar information into the model, as the factored models, but directly without adding additional model features. However, this approach may result in sparser models.

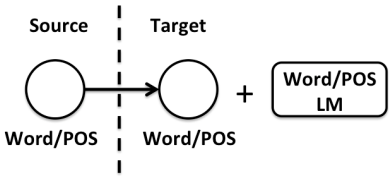

Figure 2. Translation in PB-SMT-POS

To illustrate the differences between the two baselines we show example phrase pairs taken from the respective phrase tables, below.

Example phrase pairs (PB-SMT baseline):

		cafe		
		cafeteria		
		coffee , please ?		

Example phrase pairs(PB-SMT-POS baseline):

/NN.Food ||| cafe/NN |||
/NN.Food ||| cafeteria/NN |||
/NN.Food ||| coffee/NN ,/, please/UH |||

Table 1 is populated with the results from the PB-SMT and PB-SMT-POS baselines. Although the BLEU scores of the PB-SMT-POS baseline on Myanmar to Japanese and Japanese to Myanmar are comparable to those from the PB-SMT baseline, they are considerably lower for Myanmar to English and English to Myanmar translations (-0.92 BLEU score for my-en and -0.37 BLEU score for en-my). We will discuss this in the next section.

Table 1. BLEU scores of two baselines

Source-Target	PB-SMT	PB-SMT-POS
my-ja	27.74	27.64
ja-my	24.70	24.66
my-en	21.11	20.19
en-my	23.25	22.88

5.2. Factored Translation Configurations

Our experimentscan be divided into two groups: experiments inthe first group usedonly one translation path, and experiments in the second group used two translation paths between a subset of factors. We employ a similar notation as in [2]to refer to factored translation models as follows:

"t" stands for "translation"
"g" stands for "generation"
"W" stands for "surface or word"
"P" stands for "Part-of-Speech (POS)"
"+" stands for "adding one more factor"

For example, tW-W is a translation factor for "word" to "word" translation; tW-WP is a translation factor for "words" to "words and POS tags"; tW-W+tP-P is a translation factor that uses two translation paths, one is W-W (words to words) and another is P-P (POS tags to POS tags). In this paper, we experimented with fourdifferent translation factors in thefirst group (tW-W, tW-WP and tWP-W, tWP-WP) and eight different translation factors (tW-W+tP-P, tWP-W+tP-P, tW-WP+tP-P, tWP-WP+tP-P, tW-W+tWP-W, tW-W+tWP-WP, tWP-WP+tW-W andtWP-W+tW-W) in thesecond group.In addition, we also used generation factors gW-P for creating generation tables between target factors for all of the factored experiments. Note that this generation mapping is between target factors, not between source and target factors [15]. In addition, all factored models contained a language model over the generated POS tag sequence. Therefore, the factored model "tW-W" is identical to the baseline PB-SMT, except for the target POS tag generation step and the

target POS tag language model feature. Word alignment for all factored translation models was carried out on the surface form. Based on the mapping of translation factors, the translation process will be changed (see Figure 3 and Figure 4).

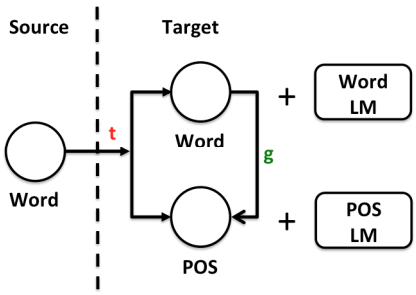

Figure 3. Translation in factored SMT with tW-WP, gW-P translation factor

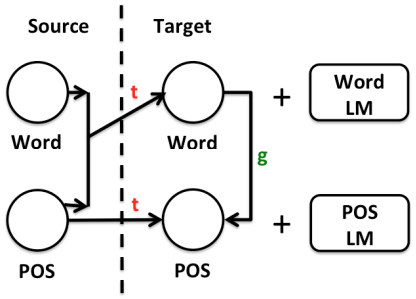

Figure 4. Translation in factored SMT with tWP-W+tP-P, gW-P translation factor

Phrase mapping and the number of phrase tables also depend on the translation factors. For example, there is only one phrase table usedto map "word" to "word|POS" in the tW-WP factored model (see Figure 3). However, there are two phrase tables, one for mapping "word|POS" to "Word" and the other for mapping "POS" to "POS in the model illustrated in Figure 4.

5.3. Results

The results displayed in Table 2 show that some factored SMT systems ofGroup 1 (using one translation path) for Myanmar to Japanese (my-ja) and Myanmar to English translation gave higher BLEU scores than both of the baselines. The translation performance of the factored models in all-but-one of the Myanmar to English experiments are significant improvements (all reported significance tests are at $p<0.05$) over thePB-SMT-POS baseline.Moreover, Myanmar to Japanese factored translation with "tW-WP" translation pathalso achieves significant improvementover both of the baselines. On the other hand, BLEU scores of translation from Japanese, English to Myanmar are lower than the PB-SMT baseline.Although BLEU scores of Japanese to Myanmar translations are lower than PB-SMT-POS, English to Myanmar translation with "tW-W", "tW-WP" and "tWP-W" translation paths can give higher BLEU scores than PB-SMT-POS.

Table 2. BLEU scores of factored SMT experiment (Group 1) for Myanmar to Japanese (my-ja), Japanese to Myanmar (ja-my), Myanmar to English (my-en) and English to Myanmar (en-my) * means BLEU comparison is significant over PB-SMT-POS with p<0.05 and ** means BLEU comparison is significant over both PB-SMT and PB-SMT-POS with p<0.05

Src-Tar	tW-W	tW-WP	tWP-W	tWP-WP
my-ja	27.73	28.17**	27.64	27.63
ja-my	24.47	24.40	24.45	24.40
my-en	20.63*	21.16*	20.70*	20.74*
en-my	22.90	22.91	23.10	22.71

Table 3. BLEU scores of factored SMT experiment (Group 2) * means BLEU comparison is significant over PB-SMT-POS and ** means BLEU comparison is significant over both PB-SMT and PB-SMT-POS

Src-Tar	tW-W +tP-P	tWP-W +tP-P	tW-WP +tP-P	tWP-WP +tP-P	tW-W +tWP-W	tW-W +tWP-WP	tWP-WP +tW-W	tWP-W +tW-W
my-ja	27.94*	27.77	28.02**	23.85	27.74	27.84	27.83	27.92**
ja-my	24.24	24.47	24.23	24.34	24.57	24.32	23.97	24.44
my-en	20.84*	20.61*	20.78*	20.88*	20.82*	20.95*	20.95*	15.54
en-my	22.48	22.96	22.92	22.94	15.55	22.91	17.17	17.46

The factored experiment results with Group 2 (using two translation paths) can be seen in Table 3. Similar tothe Group 1 results, Myanmar to English translation results with “tW-W+tP-P”, “tWP-W+tP-P”, “tW-WP+tP-P”, “tWP-WP+tP-P”, “tW-W+tWP-W”, “tW-W+tWP-WP” and “tWP-WP+tW-W” are significant better than the PB-SMT-POS baseline. Moreover, Myanmar to Japanese with “tW-W+tP-P” also gives a significant improvement. The translation performance of Myanmar to Japanese with translation paths “tW-WP+tP-P” and “tWP-W+tW-W” are significantly better than both baselines. On the other hand, the Japanese to Myanmar language translation results are lower than both of the baselines. Although the BLEU scores of English to Myanmar are lower than both of the baselines, four translation paths “tWP-W+tP-P”, “tW-WP+tP-P”, “tWP-WP+tP-P” and “tW-W+tWP-WP” gave higher BLEU scores than the PB-SMT-POS baseline.

6. Discussion

In this section, we discuss the results presented in Table 1 (two baselines),Table 2 (factored translation of Group1) and Table 3 (factored translation of Group2).

Table 2 shows the results of the simplest experiment. In column 2 of the table the factored model is “tW-W”. This is the closest factored model to the PB-SMT baseline and includes a standard word-token-based phrase-table. The factored models in columns 3-5 all include POS tags in the tokens on the source side, the target side or on both sides. We were concerned that these POS tags may have caused issues with sparseness in the model, but this does not seem to be the case. All of the experiments gave similar levels of machine translation performance, and in fact the best performing factored model appears to be “tW-WP” which beat both baselines for my-ja. Moreover, in Table 2 and Table 3, for my-jathe best two factored translation systems contain the “tW-WP” translation factor.

In Table 3, for experiments where Myanmar is the source language, there appears to be tendency for factored models with “tWP” on the source side to give low performance. For example in Myanmar to English with factors “tWP-WP”, “tWP-WP+tP-P”, and “tW-W+tWP-W”, we observed that: tW-WP > tWP-W, tW-WP > tWP-WP, tW-WP+tP-P > tWP-W+tP-P.

From the evidence above, we can believethat “tW-WP” translation path is useful feature for Myanmar to English and Myanmar to Japanese factored translation. We concludethat the POS tag factors of English and Japanese are making an effective contribution.

The results in Table 2 and Table 3 also indicate that in Japanese to Myanmar and English to Myanmar factored translation it is difficult to get an improved performance over both of the baselines. For example, the BLEU score 23.10 of English to Myanmar translation tWP-W decreases to 22.71 for tWP-WP (when adding POS tag information to target side Myanmar). Another example is, in Japanese to Myanmar translation, BLEU scores are the same for both tW-WP and tWP-WP. One possible reason is the quality of the Myanmar POS tagger. As we mentioned in Section 4.1, we used only 3000 lines of POS tagged data for building a model and the percentage of unknown tags is nearly 40% in the training data. Another important factor is that domain of the BTEC corpus is travel and the training data for POS tag modeling was from the general domain.

For two translation paths, English to Myanmar translation with “tW-W+tWP-W”, “tWP-WP+tW-W” and “tWP-W+tW-W” give very low BLEU scores compared to other translation paths. Currently we are unsure of the causes of this degradation in performance. The explanation remains future research.

There seems a tendency for translations between Myanmar and Japanese to be improved by adding a second generation step that generates POS tags or Word/POS tags from the same. Whereas in the English/Myanmar translation experiments, adding this second generation step tended to degrade performance. We believe the explanation lies in the similarity of the Myanmar language to Japanese. Both languages are syllabic, and often word tokens in both languages have similar grammatical functions (for example the Japanese function word “が” (ga) corresponds to the Myanmar function work “□” (ka), whereas there is no such word in the English language). Therefore, due to the more direct correspondence, we believe that mappings involving POS tags between the languages, are likely to be more effective between Myanmar and Japanese than with English and Japanese.

Although factored SMT can give higher BLEU scores for Myanmar to Japanese and Myanmar to English translation, decoding (and therefore tuning)takes 2 to 5 times longer than with typical PB-SMT. The computational expense was especially high when two translation paths approaches were used.

7. Conclusion

The experimental results show that factored machine translation for the low-resource language Myanmar gave higher translation performance than typical PB-SMT. However, care must be taken in choosing the type of factored model used. Due to the lack of training data for the POS tag factor, the Myanmar annotations for this factor tended to be incomplete, and potentially inaccurate (we assume this, but did not measure it). We found that using the Myanmar POS tag factor on the source side, did not seem degrade performance, but did degrade performance when used on the target side. On the other hand, we found that using factors involving POS-tags on the target side where the target was a high-resource language, could result in improved performance.

We plan to extend our study on factored machine translation in the futurewithhigher quality POS tagged Myanmar data. Based on our experimental results, we are optimistic that for similar languages, such as Japanese, factored models will become effective as the quality of the annotation of the factors improves.

Acknowledgement

We would like to thank University of Computer Studies, Yangon (UCSY) for sharing their POS tagged Myanmar data for this experiment.

References

[1]Philipp Koehn and Hieu Hoang, "Factored Translation Models", *In Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, ACL, Prague, June 2007, pp. 868-876.

[2]OndrejBojar, "English-to-Czech Factored Machine Translation", *In Proc. of the 2nd Workshop on SMT*, ACL, Prague, June 2007, pp.232-239

[3] OndrejBojar and Jan Hajic, "Phrase-Based and Deep Syntactic English-to-Czech Statistical Machine Translation", *In Proc. of the 3rd Workshop on SMT*, ACL, Columbus, Ohio, USA,June 2008, pp.143-146

[4] Helena de Medeiros Caseli and Israel AonoNunes, "Factored Translation between Brazilian Portuguese and English", *In Proc. of the 20th Brazilian Conference on Advances in Artificial Intelligence (SBIA 2010)*, Bernardo do Campo, Brazil, 2010, pp.163-172

[5]StephaneHuet, Elena Manishina and FabriceLef ´ evre, "Factored Machine Translation Systems for Russian-English", *In Proc. of the 8th Workshop on SMT, ACL, Sofia*, BulgariaAugust 2013, pp.154-157

[6]GenichiroKikui, Seiichi Yamamoto, Toshiyuki Takezawa, and EiichiroSumita (2006), "Comparative study on corpora for speech translation",*In IEEE Transactions on Audio, Speech and Language*, 14(5), pp.1674-1682

[7] Helmut Schmid, "Probablistic Part-of-Speech Tagging Using Decision Trees", *In Proc. of International Conference on New Methods in Language Processing*, Manchester, UK,1994, pp. 44-49

[8] TakuKudo, Kaoru Yamamoto and Yuji Matsumoto, "Applying Conditional Random Fields to Japanese Morphological Analysis", *In Proc. of the EMNLP*, 2004, pp. 230-237

[9]PhyuHninnMyint, Tin MyatHtwe and Ni LarThein, "Bigram Part-of-Speech for Myanmar Language", *In Proc. of theICICM*, Singapore, October 2011, pp.147-152

[10] Yuan Liu, Qiang Tan, and Kun XuShen,"The Word Segmentation Methods for Chinese Information Processing (in Chinese)",QuingHua University Press and Guang Xi Science and Technology Press, 1994, Page 36.

[11] Pak-kwong Wong and Chorkin Chan, "Chinese Word Segmentation based on Maximum Matching and Word Binding Force", *In Proc. of the 16th Conference on Computational Linguistics, Volume 1*,Copenhagen, Denmark, 1996, pp. 200-203

[12] Graham Neubig, Yosuke Nakata, Shinsuke Mori, "Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis", *In Proc. of the 49th ACL-HLT*, short papers Volume2, Portland, Oregon, USA, June 2011,pp. 529-533

[13] Marcello Federico and Mauro Cettolo, "Efficient Handling of N-gram Language Models for Statistical Machine Translation", *In Proc. of the 2nd Workshop on SMT*, Prague, Czech, 2007, pp. 88–95

[14] Franz Och and Hermann Ney, "Improved Statistical Alignment Models", *In Proc. of ACL*, Hong Kong, October 2000, pp. 440-447

[15] MOSES version 0.91, A Factored Phrase-based Beam-search Decoder for Machine Translation.

URL: http://www.statmt.org/moses/

[16] Qin Gao and Stephan Vogel, "Parallel implementations of word alignment tool", *In Proc. of the ACL Workshop: Software Engineering, Testing,*

and Quality Assurance for Natural Language Processing, Columbus Ohio, USA, 2008, pp. 49-57

[17] Kishore Papineni, SalimRoukos, Todd Ward and Wei Jing Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation", *In Proc. of the 40th ACL*, Philadelphia, USA, 2002, pp. 311-318

[18] Philipp Koehn, "Statistical Significance Tests for Machine Translation Evaluation", *In Proc. of the EMNLP*, Barcelona, Spain, 2004, pp. 388-395

[19] Evaluation tool mteval-v13a.pl
URL: http://www.itl.nist.gov/iad/mig/tests/mt/2009/