Myanmar Part of Speech Tagging Based on Machine Translation

Phyo Thu Htet
Department of Information Science
University of Technology (Yatanarpon Cyber City)
Pyin Oo Lwin, Myanmar
phyothuhtet@utycc.edu.mm

Naing Linn Phyo
Department of Information Science
University of Technology (Yatanarpon Cyber City)
Pyin Oo Lwin, Myanmar
nainglinnphyo@utycc.edu.mm

Thiha Nyein
Department of Information Science
University of Technology (Yatanarpon Cyber City)
Pyin Oo Lwin, Myanmar
thihanyein@utycc.edu.mm

***Abstract*—This paper compares the performances achieved by Phrase-Based Statistical Machine Translation system (PBSMT), Neural Machine Translation based on sequence model of LSTM (Long Short Term Memory), and Attention-Based Neural Machine Translation systems when translating Part-Of-Speech tagging (POS tagging) in Myanmar Language. Part-Of-Speech tagging is essential in natural language processing: most NLP tasks such as building lemmatizers, text classification, spelling checkers, and others require POS Tagging as a foundation step. The three approaches use parallel data Myanmar-POS tagging corpus with word segmented 10k text. Through the research, PBSMT outperforms NMT. PBSMT also has a much higher evaluation score in comparison. The evaluation process of this research uses metrics RIBES score, BLEU score, and ChrF^{++}. The result shows the Statistical approach outperforms Neural Machine Translation based on the data used in this research.**

***Index Terms*—NLP, POS Tagging, Machine Translation, PBSMT, LSTM, Attention**

I. Introduction

Part of speech is a category to which a word is assigned by its syntactic functions and so the POS tagging is the activity of marking up a word in a text (corpus) as corresponding to a particular part of speech. Part of Speech tagging (also known as Grammatical tagging) is one of the most important NLP research processes. POS also is an important port in Linguistics. As it has a relationship with adjacent and related words in a phrase, sentence, or paragraph, Part-Of-Speech Tagging stands as a fundamental role in NLP research processes. Myanmar is one of the most spoken languages in Myanmar by the people in plain and also a sub-language for most people of the hills and belongs to the subfamily of Sino-Tibetan languages.

Myanmar is one of the under-resourced languages and there is no exact rule for word boundaries. Myanmar Language also needs to be POS tagged(Grammatical tagging). Like other translations (E.g. English to Spanish), Myanmar and POS tags can use the machine translation approach. Myanmar sentence is input as source language and POS would be the target language. For example,
Input: ဦးသန့် အား ၎င်း မွေးဖွား ရာ ပန်းတနော် မြို့ ကို ရည်စူး ၍ ပန်းတနော် ဦးသန့် ဟု အမည်တွင် ခဲ့ သည် ။
Output: n ppm pron v part n n ppm v conj n n part v part ppm punc

In the output, “n” refers to “noun”, “ppm” means “pronoun”, “part” stands for “particle”, “v” means “verb”, etc. Neural Machine Translation, especially Attention-based models, and Phrase-based Machine Translation both recently become the choices among translate methods. In Neural Machine Translation, translation is performed by the encoder, decoder, and hidden layers. In this paper, we focused on both the Sequence to Sequence Model and also Sequence to Sequence with an Attention mechanism. Statistical methods have normally based on many approaches but we choose to perform with a phrase-based mechanism. Phrase tables mapped one-to-one in parallel.

II. Related Work

There were may works that tried to solve the problem of POS Tagging. [1] This work compares the performances achieved by Phrase-Based Statistical Machine Translation systems (PBSMT) and attention-based Neural Machine Translation systems (NMT) when translating User Generated Content (UGC) from French to English. It provide that PBSMT has overcomerd NMT in performance. In this paper, a new approach to the Part-of-Speech (PoS) tagging problem with machine translation. The phrases convey contextual information that can be very useful in the PoS tagging problem. Phrase-based statistical machine translation is used as an approach [2]. Neural Machine Translation (NMT) plays an important role in current natural language processing (NLP) community and its performance is usually used as a metric to evaluate the development of artificial intelligence. Machine translation has become an irreplaceable application in the use of mobile phones. In this paper, improve the performance of neural machine translation (NMT) with shallow syntax

(e.g., POS tag) of target language, which has better accuracy and latency than deep syntax such as dependency parsing [3].

III. Data

The data texts used are word-segmented. In this research, there are 15 part-of-speech tags for Myanmar Language. The Part-of-Speech tags used in datasets are abb, adj, adv, conj, fw, int, n, num, part, ppm, pron, punc, sb, tn, and v. The data and also the tag definition are used from the previous work of Dr. Ye Kyaw Thu (https://github.com/ye-kyaw-thu/myPOS).

A. Tags

The meanings and examples of the tags can be described as
1. abb as Abbreviation (E.g. အ.မ.က, အ.လ.က),
2. adj as Adjective (E.g. ဝသော, ပိန်သော),
3. adv as Adverb (E.g. နှေးနှေး, ဖြေးဖြေး),
4. conj as Conjunction (E.g. နှင့်, ဖြင့်),
5. fw as Foreign Word (E.g. 1, 2, Facebook),
6. int as Interjection (E.g. အောင်မလေး),
7. n as Noun (E.g. ကား, နွား), 8. num as Number (E.g. ၁, ၂, ၃),
9. part as Particle (E.g. ပြီး, များ),
10. ppm as Post-positional Marker(E.g. သည်, က, ကို, သို့, မှာ, တွင်),
11. pron as Pronoun(E.g. ကျွန်တော်, ကျွန်မ, သင်, သူ),
12. punc as Punctuation(E.g. ။, ၊),
13. sb as Symbol(E.g. %, $),
14. tn as Text Number (E.g. တစ်, နှစ်, ငါး)
15. v as Verb (E.g. လှုပ်ရှား, သွားလာ)

B. Data Preprocessing

Data Formatting is performed.
Original Data: ယခုn လn တွင်ppm ပျားရည်n နှင့်conj ပျားဖယောင်းn များpart ကိုppm စုဆောင်းv ကြpart သည်ppm ဟုpart ခန့်မှန်းv နိုင်part သည်ppm ။punc
Data: ယခု လ တွင် ပျားရည် နှင့် ပျားဖယောင်း များ ကို စုဆောင်း ကြ သည် ဟု ခန့်မှန်း နိုင် သည် ။<|||>n n ppm n conj n part ppm v part ppm part v part ppm punc
Data is shuffled and divided into training, development, and testing. In neural machine translation, "start" and "end" are ntroduced in the target sentences. E.g. <start> n ppm n n ppm ppm n adj n adj v part part ppm punc <end> Formatting to fit into pandas dataframe is executed. The preprocessing stages of word indexing, sentence indexing, and padding are performed in LSTM, and LSTM with attention. Perl, python, and shell programs are used.

IV. Methodologies

This section will be used to deliver the methodologies used in this research.

A. Phrase-based Statistical Machine Translation (PBSMT)

Statistical Machine Translation is a machine translation paradigm where the translation is generated based on statistical models whose parameters are derived from the analysis of bilingual text corpora. In this paper, the Phrase-based machine translation is used to translate the Myanmar word segmented sentence to POS Tag sentence. A PBSMT translation model is based on phrasal units [4]. Here, a phrase is simply a contiguous sequence of words and generally, not a linguistically motivated phrase. A phrase-based translation model typically gives better translation performance than word-based models. We can describe a simple phrase-based translation model consisting of phrase-pair probabilities extracted from the corpus and a basic reordering model, and an algorithm to extract the phrases to build a phrase-table [5].

The phrase translation model is based on a noisy channel model. To find best translation $\hat{e}$ that maximizes the translation probability $\mathbf{P}(f)$ given the source sentences; mathematically. The translation is modeled as equation

$$\hat{e} = argmax_e \mathbf{P}(e|f) \tag{1}$$

Applying the Bayes' rule, we can factor it into three parts.

$$P(e|f) = \frac{\mathbf{P}(e)}{\mathbf{P}(f)} \mathbf{P}(f|e) \tag{2}$$

The final mathematical formulation of the phrase-based model is as follows:

$$argmax_e \mathbf{P}(e|f) = argmax_e \mathbf{P}(f|e)\mathbf{P}(e) \tag{3}$$

We note that denominator $\mathbf{P}(f)$ can be dropped because for all translations the probability of the source sentence remains the same. The $\mathbf{P}(e|f)$ variable can be viewed as the bilingual dictionary with probabilities attached to each entry to the dictionary (phrase table). The $\mathbf{P}(e)$ variable governs the grammaticality of the translation and we model it using the n-gram language model under the PBMT paradigm.

B. LSTM Long Short-Term Memory (LSTM)

Neural machine translation is an approach to machine translation that uses the neural network to predict the likelihood of a sequence of words, typically modeling entire sentences in a single integrated model.LSTM (https://www.researchgate.net/publication/13853244$_{L}ong_{S}hort-term_{M}emory$) networks are a type of recurrent neural network capable of learning order dependence in sequence prediction problems.

$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f)$
$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i)$
$o_t = \sigma_c(W_o x_t + U_o h_{t-1} + b_o)$
$\tilde{c}_t = \sigma_c(W_c x_t + U_c h_{t-1} + b_c)$
$c_t = f_t \circ c_{t-1} + i_t \circ \tilde{c}_t$
$h_t = o_t \circ \sigma_h(c_t)$

where the initial values are $c_0 = 0$ and $h_0 = 0$ and the operator $\circ$ denotes the Hadamard product (element-wise product). The subscript t indexes the time step. Variables

$x_t \in \mathbb{R}^d$: input vector to the LSTM unit $f_t \in \mathbb{R}^h$: forget gate's activation vector $i_t \in \mathbb{R}^h$: input/update gate's

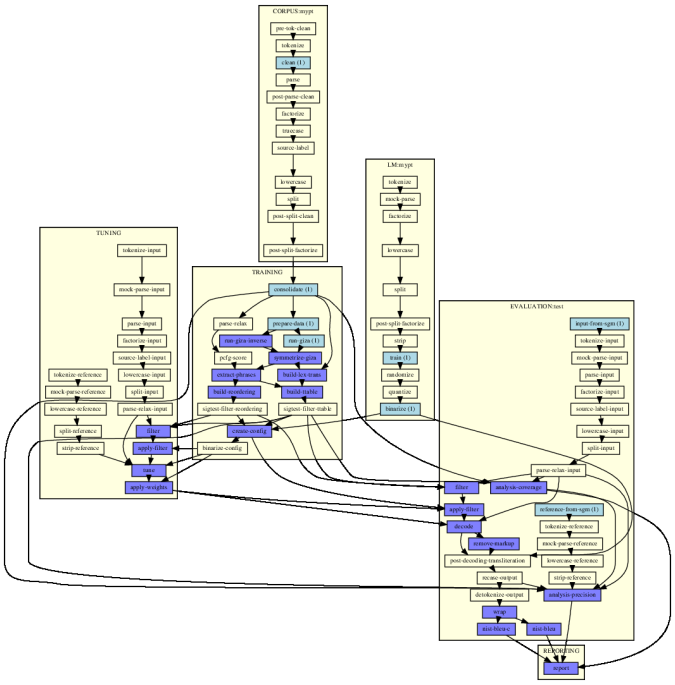

Fig. 1. The flow of PBSMT approach used in this paper

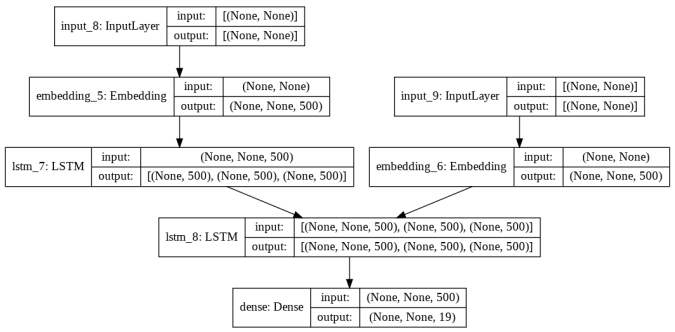

Fig. 2. Architecute of LSTM (Encoder-Decoder) Model

activation vector $o_t \in \mathbb{R}^h$: output gate's activation vector $h_t \in \mathbb{R}^h$: hidden state vector also known as output vector of the LSTM unit $\tilde{c}_t \in \mathbb{R}^h$: cell input activation vector $c_t \in \mathbb{R}^h$: cell state vector $W \in \mathbb{R}^{h \times d}$,$U \in \mathbb{R}^{h \times h}$ and $b \in \mathbb{R}^h$: weight matrices and bias vector parameters which need to be learned during training

where the superscripts d and h refer to the number of input features and number of hidden units, respectively. In the encoder model, there is an Embedding layer after the input layer, and then An LSTM layer. LSTM layer produced encoder output and two encoder stage output, which will be the input of another LSTM layer. In the decoder model, there is a Decoder input layer followed by the Decoder Embedding Layer. The output stage of the encoder is the input to the decoder LSTM layer. Finally, the dense layer with softmax activation function is used to produce the decoder output.

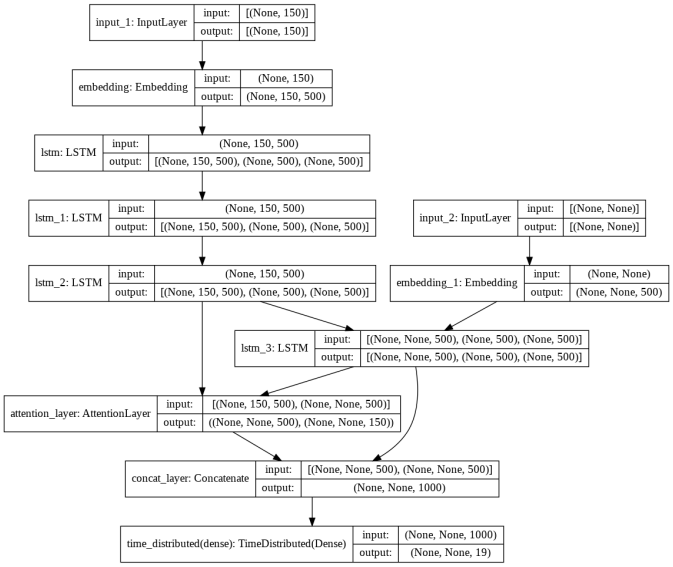

Fig. 3. Architecute of LSTM with Attention Mechanism

C. LSTM with Attention Mechanism

The major drawback of encoder-decoder models in sequence to sequence recurrent neural networks is that it can only work on short sequences. It is difficult for the encoder model to memorize long sequences and convert it into a fixed-length vector. Moreover, the decoder receives only one information that is the last encoder hidden state. Hence it's difficult for the decoder to summarize large input sequences at once. So, how do we overcome this problem? Now, this is where the concept of 'Attention Mechanism' comes. The major intuition about this is that it predicts the next word by concentrating on a few relevant parts of the sequence rather than looking on the entire sequence. There are mainly two types of attention mechanism:Global Attention and Local Attention [6]. Global Attention: Global Attention is those attention in which all the hidden state vectors of the encoder are passed to get the context vector. Local Attention:Local Attention is those attention in which only a few hidden state vectors of encoder are considered for the generation of context vectors. Global Attention called Bahdanau Attention is used in this research.
$score(h_t, \bar{h}_s) = v_a^\top tanh(W_1 h_t + W_2 \bar{h}_s)$[Bahdanau's additive style]
Score = FC(tanh(FC(EO)+FC(H)))

$\alpha_{ts} = \frac{exp(score(h_t, \bar{h}_s))}{\sum_{s\prime=1}^{S} exp(score(h_t, \bar{h}_{s\prime}))}$[Attention weights]
Attention weights = softmax(score, axis =1)
$c_t = \sum_s \alpha_{ts} \bar{h}_s$[Context vector]
$a_t = f(c_t, h_t) = tanh(W_c[c_t; h_t])$[Attention vector]
Attention vector = concat(embedding out,context vector)

In this experiment, the global attention layer is used between the encoder output and the decoder input. All the encoder output is feed into the attention layer and the output of the attention layer is the first decoder input.

V. Experimental Setup

A. Machine Translation

Statistical Machine Translation for POS tagging is performed by using Moses [7] enforced with Giza^{++} [8] for word alignment of the parallel corpus. The KenLM language Model is utilized as the language model of POS Tagging. The framework called Tensorflow is used for neural machine translation for both LSTM and LSTM mechanisms. The version of Tensorflow is 2.0, and Google Colab is used to acquire GPU resources in training these models or architectures. NLTK library for BLEU Score and Word Error Rate are also used.

B. Evaluation

To evaluate our experiment, BLEU (bilingual evaluation understudy) [9]. BLEU: a method for automatic evaluation of machine translation. RIBES (Rank-based Intuitive Bilingual Evaluation Score [10] and Chrf++(character n-gram F score) [([https://github.com/m-popovic/chrF])] is used. chrF++ is a tool for automatic evaluation of machine translation output based on character n-gram precision and recall enhanced with word n-grams. The tool calculates the F-score averaged on all character and word n-grams, with the default character n-gram order of 6 and word n-gram order of 2. In calculating BLEU score, 4gram is used with weights = (0,0,0,1). In RIBSE score, alpha (0.25) is used

VI. Results and Discussion

The evaluation is performed on the test data of 1099 sentences. Table I illustrates the evaluation scores on three different metrics (BLEU, RIBES, ChrF^{++}) of the closed test. As of Table I, the PBSMT approach provides the results of BLEU score (0.7727), RIBES score (0.9726), and ChrF^{++}(89.05). LSTM Model achieves BLEU score (0.4529), RIBES score (0.8646), and ChrF^{++}(75.25). The results of LSTM with attention mechanism are BLEU score (0.7699), RIBES (0.9659), and finally ChrF++ (85.57).

Overall, PBSMT provides the best result for the BLEU score. On the other hand, the lowest score results in LSTM Model. In terms of the RIBES score. For the RIBES score, the highest RIBES score is obtained by using PBSMT. LSTM with the global attention mechanism of Bahdanau Attention performs the best in ChrF^{++} evaluation.

The performance of PBSMT in all of the categories of BLEU, RIBES, and ChrF^{++} is the best. LSTM enhanced with attention Mechanism is also working well in these scores. For the current data, the architecture of LSTM can not perform well regarding BLEU Score.

VII. Error Analysis

In PBSMT, most of the errors are names, numbers that not include in data, foreign words, and single words. PBSMT can not predict the OOV (out of vocab) words. The following one can be described as an example.

Input: ၁၉၄၂ ခုနှစ် တွင် ရန်ကုန် မြို့ စမ်းချောင်း မြို့နယ် မှ (ဒေါ်နော်မာ) နှင့် အိမ်ထောင်ကျ ကာ (ဒေါ်ဦးဦးမေ) ၊ (ဦးကျော်ကျော်) ၊ (ဒေါက်တာ အောင်သော်) ၊ (ဒေါ်မော်မော်) နှင့် (ဒေါ်သန်းသန်းဆွေ) စသည့် သားသမီး များ ထွန်းကား ခဲ့ သည် ။

Output: num n ppm n n n n ppm ppm v conj (ဒေါ်နော်မာ) (ဒေါ်ဦးဦးမေ) punc (ဦးကျော်ကျော်) punc n (အောင်သော်) punc (ဒေါ်မော်မော်) conj (ဒေါ်သန်းသန်းဆွေ) part n part v part ppm punc

For Sequence to Sequence model of LSTM, the confusion pairs are especially part ==> n, v ==> n, n ==> part. LSTM can work well on short sentences but not on long ones.

In LSTM with attention mechanism, most of the OOV are predicted as "n" which is the maximum number of tags in the training data. For example,

Original Input: ၁၈၆၀ ခုနှစ် တွင် (ဒီလရှယ်လီဘရားသားစ်) က ခရစ်ယာန် သာသနာပြု ကျောင်း များ ကို တည်ဆောက် ခဲ့ ကြ သည် ။

Input: ၁၈၆၀ ခုနှစ် တွင် (OOV) က ခရစ်ယာန် သာသနာပြု ကျောင်း များ ကို တည်ဆောက် ခဲ့ ကြ သည် ။

Output: num n ppm (n) ppm n n n part ppm v part part ppm punc

Since POS tagging is based on the approach of Machine Translation, the predicted length is not the same as original length of the sentence. And he maximum length of predicted POS sentences is 25.

Input: သို့နှင့် မဂ္ဂင်း မှ တစ်ဆင့် သတင်းစာ ကို ပါ တိုးချဲ့ လိုက် သောအခါ တွင် ဘက်ပတစ် ကျောင်း သို့ မ ပြန် တော့ ဘဲ ထို မဂ္ဂင်း ၊ သတင်းစာ နှစ် ခု စလုံး တွင် ပင် တည်းဖြတ် သည့် ဘက် မှ ဆက်လက် လုပ်ကိုင် လေ တော့ သည် ။ (Length: 37)

Output: conj n ppm adv n ppm part v part conj ppm n n ppm part v part part adj n punc n tn part part (Length: 25)

VIII. Conclusion

POS Tagging is fundamental that is crucial in other Natural Language Processing tasks. According to the results, Myanmar POS Tagging based on Machine Translation is also useful by giving the appropriate performance. Because of the low amount of data, Phrase-Based Statistical Machine Translation provides better results than the other approaches used in this research.

Neural Machine Translation with the attention mechanism can also provide excellent results (96.33% in RIBES score). With more data, the results can be different. Since Myanmar is an under-resourced language, Statistical approaches will be more useful for the current situation. Overall, the machine translation approach to perform POS tagging would be a potential alternative.

In the future, the process of hyperparameter tuning will also perform. The research will also extend by using the metrics like Accuracy, Precision, Recall, and so-on. More data will introduce to the current data. Supplementing with the segmentation schemes like character-level segmentation, syllable level segmentation, and sub-word level segmentation of a sentence would be interesting.

Acknowledgment

The deepest gratitude would like to be delivered to Dr Ye Kyaw Thu and Dr Hnin Aye Thant for their guidances concerning this research.

TABLE I
BLEU,RIBES AND CHRF^{++} SCORES FOR PBSMT, LSTM, LSTM WITH ATTENTION OF MYANMAR POS TAGGING (BOLD NUMBERS INDICATED THE HIGHEST SCORES)

	BLEU	**RIBES**	**ChrF^{++}(c6+w2-avgF)**
PBSMT	**0.7727**	**0.9726**	**89.05**
LSTM	0.4529	0.8647	75.25
LSTM with Attention	0.7699	0.9659	85.57

REFERENCES

[1] Jos□Carlos Rosales Nez, "A Comparison between NMT and PBSMT Performance for Translating Noisy User-Generated Content", Universit□ Paris Sud, LIMSI.

[2] Guillem Gasco i Mora and Joan Andreu Sanchez Peiro, "Part-of-Speech Tagging Based on Machine Translation Techniques", Departament de Sistemes Inform`atics i Computacio Universitat Polit`encia de Val`encia Cam□de Vera s/n, 46022 Val`encia (Spain).

[3] Xiaocheng Feng, "Enhanced Neural Machine Translation by Joint Decoding with Word and POS-tagging Sequences", Springer Science+Business Media, LLC, part of Springer Nature 2020.

[4] Koehn, Philipp, and Och, Franz Josef and Marcu, Daniel, "Stastatistical phrase-based translation," Proceedings of the 2003 Con- ference of the North American Chapter of the Association for Computational Linguistics on Human Language TechnologyVolume 1, 2003, pp. 48–54.

[5] Lucia Specia□"Tutorial, Fundamental and New Approachesto Statistical Machine Translation," International Conference.

[6] D. Kauchak and R. Barzilay, ``Paraphrasing for automatic evaluation,'' Asso. Com. Ling. New York City, USA, vol. Pro. Hum. Lang. Tech. Conf. NAACL, Main Conference, pp. 455––462, June 2006.

[7] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin and E. Herbst, ``Moses: Open source toolkit for statistical machine translation,'' Asso. Com. Linguistics, booktitle. Proc. ACL-08: HLT, Proc. 45th Ann. Meet. Asso. Com. Ling. Comp. vol. Proc. Demo and Poster Sessions, Prague, Czech Republic, pp. 177––180, June 2007.

[8] F. Braune, A. Gojun and A. Fraser, ``Long-distance reordering during search for hierarchical phrase-based SMT,'' In Proc. of the 16th Ann. Conf. of the Euro. Asso. for Mac. Translation, Trento, Italy, pp. 177––184. 2012.

[9] Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. J. (2002).

[10] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, Hajime Tsukada Automatic Evaluation of Translation Quality for Distant Language Pairs