String to Tree and Tree to String Statistical Machine Translation for Myanmar Language

[†]Ye Kyaw Thu, [†]Andrew Finch, [‡]Win Pa Pa, [§]Khin War War Htike, [†]Eiichiro Sumita
[†]*Multilingual Translation Lab., NICT, Kyoto, Japan*
[‡]*Natural Language Processing Lab., UCSY, Yangon, Myanmar*
[§]*School of Information Science and Engineering, Central South University, Changsha, China*
yekyawthu, andrew.finch, eiichiro.sumita@nict.go.jp,
winpapa@ucsy.edu.mm, khinwarwarhtike@csu.edu.cn

Abstract

This paper contributes the first published evaluation of the quality of string-to-tree (S2T) and tree-to-string (T2S) statistical machine translation methods between Myanmar and Chinese, English, French, German in both directions. The performance of machine translation was automatically measured in terms of BLEU and RIBES scores for all experiments. In addition we performed a comparative study of the performance of phrase-based statistical machine translation (PBSMT) and T2S using human judgment. We found that the results obtained using the BLEU automatic evaluation metric were misleading and found that the T2S approach is suitable for distant languages to Myanmar machine translation.

1. Introduction

Our main motivation for this research is to investigate string-to-tree (S2T) machine and tree-to-string (T2S) translation performance for Myanmar language. Eight language pairs (Myanmar to Chinese, Myanmar to English, Myanmar to French, Myanmar to German, Chinese to Myanmar, English to Myanmar, French to Myanmar, and German to Myanmar) were used in the experiments, and translation quality was evaluated using both the BLEU and RIBES evaluation metrics. We trained the baseline PBSMT, S2T and T2S machine translation systems using a parallel corpus of 161,882 sentence pairs for each language pair.

2. Related Work

In recent years, researchers have explored various approaches to incorporate syntax and structure into statistical machine translation models. The S2T translation model is able to exploit features from source strings and target trees [2, 3, 4], T2S translation model uses source trees and target strings [5, 6, 7] and tree-to-tree (T2T) translation models exploit both source and target trees [8, 9, 10]. As far as the authors are aware there has been no evaluation of T2S, S2T and T2T statistical machine translation for the Myanmar language. Currently there is no publicly available parser for the Myanmar language and thus, it was impossible to study the T2T approach here.

3. Methodology

3.1. PBSMT

Current SMT systems are based on the translation of phrases rather than word-by-word translation. Therefore, the translation model is based on phrasal units. Here, a phrase is simply a contiguous sequence of words and generally, not linguistically motivated phrase. A phrase-based translation model gives better translation performance than word-based models.

We can describe a simple phrase-based translation model consisting of phrase probabilities extracted from corpus and a basic reordering model, and how to extract the phrases to build a phrase-table [12].

The mathematical formulation for phrase-based models is the same as that for word-based models as follow:

$$argmax_{\boldsymbol{e}}\, P(\boldsymbol{e}|\boldsymbol{f}) = argmax_{\boldsymbol{e}}\, P(\boldsymbol{f}|\boldsymbol{e})P(\boldsymbol{e})$$

However, for the phrase-based models, $P(\boldsymbol{f}|\boldsymbol{e})$ is further decomposed into:

$$P(f_1^{-I}|e_1^{-I}) = \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i) d(start_i - end_{i-1} - 1)$$

Where the source sentence $\boldsymbol{f}$ is broken up into $\boldsymbol{I}$ phrases $\bar{f}_i$, each $\bar{f}_i$ is translated into a target phrase $\bar{e}_i$ and the phrase probability is represented by $\phi(\bar{f}_i|\bar{e}_i)$. In the formula, $d(start_i - end_{i-1} - 1)$ represents a distance-based reordering model.

The calculation of distortion distance is shown in Figure 1. For example, consider the second phrase pair. It starts at position 2, the previous phrase pair ends at position 5, therefore the distortion d=2-5-1=-4. For a complete description of this calculation, the reader is referred to [13] For ease of understanding, the example here follows the format used in [13].

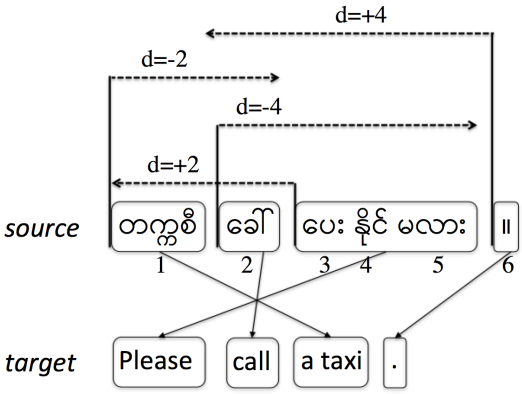

Figure 1. Calculation of distortion

3.2. Syntax-based Machine Translation

Syntax-based machine translation model uses a grammar consisting of SCFG (Synchronous Context-Free Grammar) rules with syntactic labels [14]. Applying linguistic syntax only on the source side is T2S, only on the target side is S2T and in both source and target sides is T2T. To get the annotation, a syntactic parser is required. Syntactic labels provide the structure of a sentence and can also indicate a relationship with other languages (as shown in Figure 2) Generally, the advantages of syntax-based translation are syntax-based reordering, better handling of function words, the facilitation of conditioning on syntactically related words, and enabling the use of a syntactic language model.

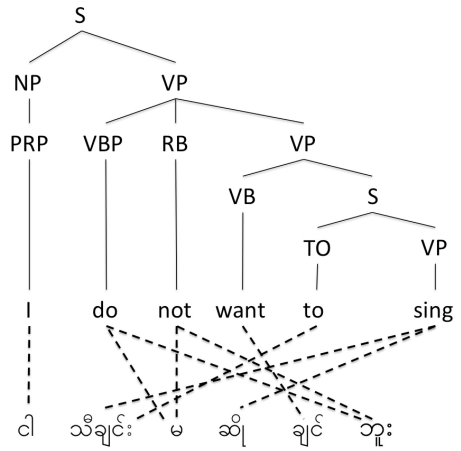

Figure 2. An example of a Myanmar sentence aligned with an English parse tree

3.2.1. String-to-Tree (ST2)

The concept of string-to-tree translation model is adding syntactic categories to target-side non-terminals in hierarchical phrase models [15]. The application of syntax-based methods (the target language parser) to annotate and generalize phrase translation tables extracted via existing phrase extraction techniques was proposed by Zollmann et al. [16]. Building translation rules that map input phrases to output tree fragments was proposed by Galley et al. [17]. Applying the EM algorithm for building contextually richer rules and learning rule probabilities that are able to lead to better translation performance was proposed in [18]. The issue of lower coverage resulting from the specificity of syntactic structure was addressed by adjusting the rule extraction algorithm (DeNeefe et al. [19]).

3.2.2. Tree-to-String (T2S)

Tree-to-string translation uses a rich source language representation for translation into word sequences in the target language. As a soft constraint, similarity between the source syntax tree and the derivation tree during decoding may be used [20], or non-terminals in rule applications that match the source syntax tree could be flagged in a feature [21]. Syntax-directed translation parses the input first and uses the parse structure to guide the output phrase generation [7]. Liu and Fildea extend this work by adding semantic role labels and changing parameter estimation [22]. Hopkins and Kuhn proposed a tree-to-string model with best-first search [23].

4. Experiments

4.1. Corpus Statistics

We used five languages from the multilingual Basic Travel Expressions Corpus (BTEC), which is a collection of travel related expressions [24]. The languages were Chinese (zh), English (en), German (de), French (fr) and Myanmar (my). 154,989 sentences were used for training, 4,999 sentences for development and 1,999 sentences for evaluation. We used Berkeley parser for tree annotation for zh, en, de and fr [25]. In all experiments, the Myanmar language was segmented using word segmentation method proposed by Win Pa Pa et al. [26].

4.2. Moses SMT system

We used the PBSMT, S2T and T2S system provided by the Moses toolkit [27] for training the PBSMT, S2T and T2S machine translation systems. The word segmented Myanmar was aligned using GIZA++ [28], and the alignment was symmetrized by grow-diag-final-and heuristic [11]. The lexicalized reordering model was trained with the msd-bidirectional-fe option [29]. We use SRILM for learning a 5-gram language model with interpolated modified

Kneser-Ney dicounting [30]. Minimum error rate training (MERT) [31] was used to tune the decoder parameters and the decoding was done using the Moses decoder (version 2.1) [27].

5. Evaluation

5.1. Automatic Evaluation

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) [32] and the other was the Rank-based Intuitive Bilingual Evaluation Measure (RIBES) [33]. 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 [32]. 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 French, Myanmar and German [33]. Large RIBES scores are better.

5.2. Human Evaluation

We used two human evaluation schemes. One was a simplified human judgment approach named "Binary system comparison" proposed by [1] and the other was a standard procedure used in manual scoring (on a 1 to 5 scale) of each sentence with two numerical values (one for fluency and one for adequacy). We performed human evaluation on PBSMT and T2S for English to Myanmar translation with four bilingual Myanmar native volunteer judges (two for each method).

5.2.1. Binary System Comparison

The main goal of this evaluation method relies in the fact that a human judge, when presented two different translations of the same sentence, can normally choose the best one in a more-or-less definite way [1]. In the social sciences, a similar method has been proposed [34]. The human judges were shown the translation output from the two systems in random order and asked to select the better translation. The judges were also shown the corresponding source sentence. No instruction on the criteria for judging the translations was given. This was for two reasons: one is to get a natural decision from judges and the other to prevent them from considering their decision explicitly in terms of fluency and adequacy [1]. An example form from the binary system comparison evaluation between PBSMT and T2S is as shown in Figure 3. Two human bilingual judges (native speakers of Myanmar) carried out the evaluation on 1,000 sentences (500 sentences for each judge).

Source: Do you have the time ?
MT System A: အချိန် ရှိ ပါသလား ။
MT System B: အခု ဘယ်နှစ် နာရီ ထိုး ပြီလဲ ။

Judgement (Select one): [A] [B] [AB]

Figure 3. Format of the binary system comparison for PBSMT and T2S (the output translations from MT System A and MT System B were randomly assigned)

5.2.2. Adequacy and Fluency

Adequacy and fluency measurement is a standard procedure for carrying out a human evaluation of machine translation output. Adequacy measurement relates to the preservation of meaning between source and

translated sentence. The fluency of a translation relates to its readability and understandability, without taking the content of the source sentence into account. We used scales that were originally developed for the annual NIST Machine Translation Evaluation Workshop by the Linguistics Data Consortium [35, 36]. The five point scale for adequacy indicates how much of the meaning expressed in the source (i.e. English) is also expressed in a hypothesis (i.e. Myanmar) translation (see Table 1). The second five point scale indicates how fluent the translation is. (see Table 1).

Table 1. Human judgment scoring scales

Adequacy		Fluency	
5	All Information	5	Flawless
4	Most Information	4	Good
3	Much Information	3	Non-native
2	Little Information	2	Disfluent
1	None	1	Incomprehensible

An example format for human judgment for adequacy and fluency of PBSMT and T2S is shown in Figures 4 and 5 respectively. Again a blind evaluation was conducted on 1,000 translated sentences from the PBSMT and T2S, systems and two human judges were used in the evaluation.

Source: Is this compartment full ?

Output: ဒီ အခန်း ပြည့် နေ ပါပြီ ။

Judgement (1-5): [ ]

Figure 4. Form for human judgment of adequacy

Output: ငွေရှင်း ကောင်တာ က ဘယ်မှာ လဲ ။

Judgement (1-5): [ ]

Figure 5. Form for human judgment of fluency

6. Results

6.1. Automatic Evaluation

The BLEU and RIBES score results for the machine translation experiments with PBSMT, ST2 and T2S are shown in Tables 2 (translating from Myanmar) and 3 (translating into Myanmar). Bold numbers indicate the highest scores of the two different approaches. Based on the results of both BLEU and RIBES scores, although we can clearly see that PBSMT gives higher translation performance than S2T (except for the my-en language pair), it is difficult to make discriminate between the results from the PBSMT and T2S approaches (see Table 3).

Table 2. BLEU and RIBES scores for translating from Myanmar

Src-Trg	PBSMT	String-to-Tree
my-de	**40.23** (**0.7735**)	37.01 (0.7685)
my-en	**44.53** (0.7878)	38.50 (**0.8140**)
my-fr	**43.01** (**0.7706**)	32.74 (0.7611)
my-zh	**35.43** (**0.7808**)	30.98 (0.7481)

Table 3. BLEU and RIBES scores for translating to Myanmar

Src-Trg	PBSMT	Tree-to-String
de-my	**39.16** (0.7571)	38.60 (**0.7778**)
en-my	**42.83** (0.7885)	40.71 (**0.8048**)
fr-my	**38.27** (0.7590)	34.02 (**0.7595**)
zh-my	29.78 (0.7051)	29.78 (**0.7149**)

6.2. Human Evaluation

This section presents the human evaluation results using four bilingual judges (all were native speakers of Myanmar). Two judges were used for binary system comparison and another two judges for the adequacy and fluency evaluation, as mentioned in Sections 5.2.1 and 5.2.2.

6.2.1 Binary System Comparison

Table 4 shows the binary system evaluation results on PBSMT and T2S. The judges were quite consistent with each other and both of them judged T2S to be better more often than PBSMT.

Table 4. Binary system comparison of results on 1000 translated sentences

Judges	PBSMT	Tree-to-String	Both
Judge-1	104	223	173
Judge-2	128	219	153

6.2.2 Adequacy and Fluency

Figure 6 shows adequacy evaluation results on PBSMT and T2S using two judges. Comparing the totals of the adequacy scores, the 2173 total score for PBSMT was comparable to the 2152 total score for T2S. Figure 7 shows fluency evaluation results on PBSMT and T2S using two judges. Comparing the totals of the fluency scores, the 2241 total score for PBSMT was lower than the 2330 total score for T2S. Moreover, both judge1 and judge2 gave a higher number of 5 ratings (flawless Myanmar) for T2S (see Figure 7).

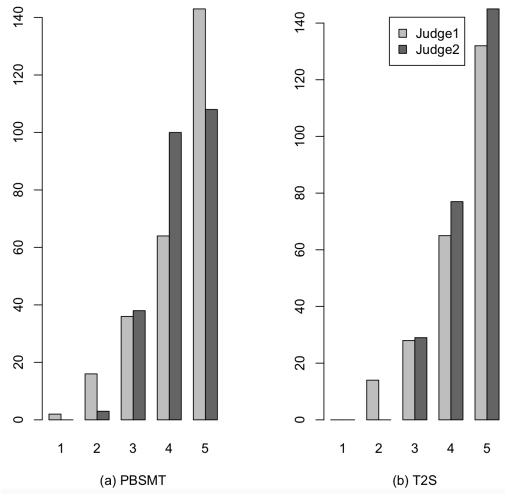

Figure 6. Evaluation score for adequacy

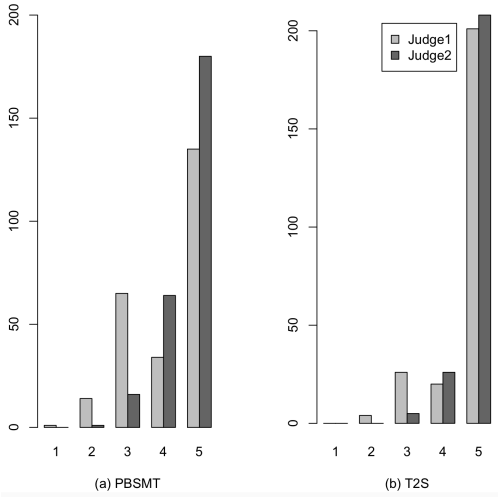

Figure 7. Evaluation score for fluency

7. Discussion

Comparing the automatic evaluation results using BLEU score for PBSMT and S2T (see Table 2), higher scores for PBSMT were observed in all experiments. Moreover higher automatic evaluation RIBES score were observed for for the my-de, my-fr and my-zh language pairs for PBSMT. This indicates that

adding syntactic categories to the target-side in the S2T approach is not effective for translation from Myanmar.

When translating to Myanmar, we got totally different result (see Table 3). In details, we can see that higher BLEU scores were observed for PBSMT when translating de-my, en-my and fr-my. On the other hand, higher RIBES scores for all four language pairs were observed for the T2S approach. Based on the automatic evaluation results of PBSMT and T2S, it is difficult to conclude which approach can provide better translation performance. Therefore, we carried out human assessments of translation quality using two human evaluation schemes (see Section 6.2.1 and 6.2.2). In this experiments, we performed human evaluation only on English to Myanmar translation output 4 bilingual Judges (all were native speakers of Myanmar).

The binary system evaluation results in Table 4 show that in 44.2% cases the T2S translation was preferred, in 23.2% of cases PBSMT was preferred and in 32.6% of cases there judges had no preference. We performed the statistical significance test proposed by D Vilar et al. [1]. The total evaluation score for a binary system comparison is as follow:

$$R_{X,Y} \coloneqq \frac{1}{m}\sum_{i=1}^{m} r_{i,X,Y}$$

Here, R is the arithmetic mean proposed by B. Efron et al. [38] for estimated standard error of the score $R_{X,Y}$, X, represented MT system X, Y and represented MT system Y and m is the number of evaluated sentences. $r_{i,X,Y}$represent a sentence score and we define as follows:

$$r_{i,X,Y} \coloneqq \begin{cases} +1 & e_{i,X}\ is\ better\ than\ e_{i,Y} \\ 0 & e_{i,X}\ is\ equal\ to\ e_{i,Y} \\ -1 & e_{i,X}\ is\ worse\ than\ e_{i,Y} \end{cases}$$

Differences between the PBSMT and T2S in the binary system evaluation were significant ($p<0.05$). Differences between the PBSMT and T2S approaches in the human adequacy evaluation were not significant, where's the results for fluency were significant (both tested at $p<0.05$). This result indicates that fluency of T2S translations was higher than those from PBSMT.

These results were surprising since the automatic evaluation results using BLEU (PBSMT>T2S) suggested that PBSMT would have higher adequacy (BLEU scores have been shown to correlate with human adequacy [32, 37]). The RIBES scores (PBSMT<T2S) were consistent with the fluency results from the human evaluation. We calculated the Kendall's Tau distance on training data for all language pairs: 0.61 (de-my), 0.53 (en-my), 0.54 (fr-my) and 0.73 (zh-my) [39]. Among them, de-my, en-my and fr-my are distant language pairs in terms of Kendall's Tau distance, which reflects the degree of word reordering required. As we mentioned in Section 5.1, the BLEU score approximately measures the adequacy of SMT and the RIBES score focuses on word order errors and is therefore suitable for evaluation of distant language pairs [33]. The RIBES score and human evaluation results, show that the BLEU scores are misleading. Moreover, when we visually inspected the translation output of PBSMT and T2S systems on en-my, we found several outputs were equal in adequacy or understandable but T2S gave better fluency. The two examples in Figure 8 show this.

Source: How long does it take for a perm ?
PBSMT: ဘယ်လောက် ကြာ မလဲ ဆံပင် ကောက် ဖို့ ။
T2S: ဆံပင် ကောက် ဖို့ အတွက် က ဘယ်လောက် ကြာ သလဲ ။

Source: Is this made in China ?
PBSMT: ဒီ ထုတ်ကုန် တရုတ် နိုင်ငံသား ပါ ။
T2S: တရုတ် နိုင်ငံ ထုတ်ကုန် လား ။

Figure 8. Two examples of similar adequacy but better fluency of T2S

On the other hand, some PBSMT outputs gave equal adequacy with T2S but better fluency (see Figure 9) but this was only approximately 3% of the test-set (1,999 sentences).

Source: Oh , happy birthday .
PBSMT: အိုး ၊ ပျော်ရွှင် မွေးနေ့ ပါ ။
T2S: မွေးနေ့ ပျော် ပါတယ် ။

Figure 9. An example of similar adequacy but better fluency of PBSMT

Compared to PBSMT translation outputs, T2S improved in reordering and approximately 75% of the test set (1,999 sentences). Ten examples of reordering mistakes in PBSMT approach compared to T2S are shown in Table 5. T2S can handle well for reordering even when out-of-vocabulary words are contained in the input sentence (for example, the first sentence of Table 5). Sentences 2 to 10 show example PBSMT reordering errors.

8. Conclusion

In this paper we have, for the first time, applied tree-to-string and string-to-tree based machine translation methods to the translation of the Myanmar language. We conducted both automatic (BLEU and RIBES), and human (binary, adequacy, fluency) evaluations on the experiments. Our result show that the tree-to-string technique was the best technique for translating into Myanmar, producing more fluent results than the standard phrase-based approach. In future work, we would like to make experiments on S2T and T2S with languages that are similar to Myanmar such as Japanese and Korean.

Acknowledgement

We would like to express our gratitude to Ms. Aye Nyein Mon and Ms. Thae Nu Htwe from Natural Language Processing Lab., University of Computer Studies, Yangon (UCSY) for their help in human evaluation.

Table 5. Reordering errors of PBSMT for English to Myanmar translation

No.	Source (English)	PBSMT Output	Tree-to-String Output
1	Have you ever seen a manta ray ?	ဖူးလား ကို manta ray ။	manta ray ကို မြင် ဖူးသလား ။
2	No sweat .	ဟင့်အင်း ချွေး ။	ချွေး မဟုတ် ပါဘူး ။
3	Do I have to change trains to go to Milan ?	ရထား တွေကို ပြောင်း စီး ကို စီး ရမှာလား မီလန် ကို သွားဘို့ အတွက် ။	မီလန် ကို သွားဘို့ အတွက် ပြောင်း မစီး လို့ မရ ဘူးလား ။
4	What is the most popular movie now ?	၊ လူကြိုက် အများဆုံး ရုပ်ရှင် က ဘာ ပါလဲ ။	အခု လူကြိုက် အများဆုံး ရုပ်ရှင် က ဘာလဲ ။
5	What 's Japan 's country code ?	ဘာ က ဂျပန် နိုင်ငံ ကုတ် နံပါတ် ။	ဂျပန် ရဲ့ နိုင်ငံ ကုတ် နံပါတ် က ဘာ ပါလဲ ။
6	Can you introduce him to me ?	ပေးနိုင်မလား ကို သူ နှင့် မိတ်ဆက် ပေးနိုင်မလား ။	ငါ့ ကို သူ နှင့် မိတ်ဆက် ပေး လို့ ရပါသလား ။
7	Is there a place I can smoke ?	လို့ ရတဲ့ နေရာ ရှိ ပါသလား ဆေးလိပ် သောက် ပါ ။	ဆေးလိပ် သောက် လို့ ရတဲ့ နေရာ ရှိ ပါသလား ။
8	Is next Saturday okay ?	လား နောက် တပတ် စနေ နေ့ ဆိုရင် အဆင်ပြေ မလား ။	နောက် တပတ် စနေ နေ့ ဆိုရင် အဆင်ပြေ မလား ။
9	Is this watch made in Switzerland ?	ဒီ နာရီ လုပ် ဆွစ်ဇာလန် နိုင်ငံ ။	ဆွစ်ဇာလန် နိုင်ငံ မှာ လုပ် နာရီ လား ။
10	Is there a guide that speaks Japanese ?	လမ်းပြ ရှိ ပါသလား ဂျပန် စကားပြော တဲ့ လူ ။	ဂျပန် စကားပြော တဲ့ ဧည့် လမ်းညွှန် များ ရှိ ပါသလား ။

References

[1] D Vilar, G Leusch, H Ney, RE Banchs. (2007). Human evaluation of machine translation through binary system comparisons, In *Proceedings of the Second Workshop on Statistical Machine Translation*, pp. 96-103.

[2] Kenji Yamada and Kevin Knight. (2001). A syntax- based statistical machine translation model. *In Proceeding of ACL 2001*, pages 132-139.

[3] Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang and Ignacio Thayer. (2006). Scalable inferences and training of context-rich syntax translation models, *In Proceeding of COLING/ACL 2006*, pp. 961-968.

[4] Libin Shen, Jinxi Xu and Ralph Weischedel. (2008). A new string-to-dependency machine translation algorithm with a target dependency language model. *In Proceeding of ACL/HLT 2008*, pp. 577-585.

[5] Chris Quirk, Arul Menezes and Colin Cherry. (2005). Dependency treelet translation: Syntactically in- formed phrasal SMT. *In Proceeding of ACL 2005*, pp. 271-279.

[6] Yang Liu, Qun Liu and Shouxun Lin. (2006). Tree-to- string alignment template for statistical machine translation. *In Proceeding of COLING/ACL 2006*, pp. 609-616.

[7] Liang Huang, Kevin Knight and Aravind Joshi. (2006). Statistical syntax-directed translation with ex- tended domain of locality. *In Proceeding of AMTA 2006*, pp. 66-73.

[8] Jason Eisner. (2003). Learning non-isomorphic tree mappings for machine translation. *In Proceeding of ACL 2003*, pp. 205-208.

[9] Yuan Ding and Martha Palmer. (2005). Machine translation using probabilistic synchronous dependency insertion grammars. *In Proceeding of ACL 2005*, Ann Arbor, Michigan, pp. 541-548.

[10] Yang Liu, Yajuan Lü and Qun Liu. 2009. Improving Tree-to-Tree Translation with Packed Forest. *In Proceeding of ACL 2009*, pp. 558-566.

[11] Koehn, P., Och, F.J., Marcu, D. (2003), Statistical phrase-based translation, *In Proceeding of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pp. 48–54.

[12] Lucia Specia, (2011). Tutorial, Fundamental and New Approaches to Statistical Machine Translation, *International Conference Recent Advances in Natural Language Processing 2011 (RANLP-2011).*

[13] Koehn, P. (2010). Statistical Machine Translation, Cambridge University Press, pp. 129-130.

[14] Syntax Tutorial of Moses (Statistical Machine Translation System), http://www.statmt.org/moses/?n=Moses.SyntaxTutorial#ntoc1

[15] Zollmann, A., Venugopal, A., Och, F.J., and Ponte, J. (2008). A systematic comparison of phrase-based, hierarchical and syntax-augmented statistical MT, *In Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1 (COLING '08)*, Vol. 1. Association for Computational Linguistics, pp. 1145-1152.

[16] Zollmann, Andreas and Ashish Venugopal (2006), Syntax augmented machine translation via chart parsing. *In Proceedings of the Workshop on Statistical Machine Translation (HLT/NAACL 2006)*, pp. 138-141.

[17] M. Galley, M. Hopkins, K. Knight, and D. Marcu. (2004). What's in a translation rule?', *in HLT-NAACL 2004: Main Proceedings, Association for Computational Linguistics*, May 2 - May 7 2004, pp. 273-280.

[18] Galley, M., Graehl, J., Knight, K., Marcu, D., DeNeefe, S., Wang, W., and Thayer, I. (2006). Scalable inference and training of ontext-rich syntactic translation models. *In Proceeding of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pp. 961-968.

[19] DeNeefe, S., Knight, K., Wang, W., and Marcu, D. (2007). What can syntax-based MT learn from phrase-based MT?. *In Proceeding of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 755-763.

[20] Zhou, B., Xiang, B., Zhu, X., and Gao, Y. (2008). Prior derivation models for formally syntax-based translation using linguistically syntactic parsing and tree kernels. *In Proceedings of the ACL-08: HLT Second Workshop on Syntax and Structure in Statistical Translation (SSST-2)*, pp. 19-27.

[21] Marton, Y. and Resnik, P. (2008). Soft syntactic constraints for hierarchical phrased-based translation, *In Proceedings of ACL-08: HLT*, pp. 1003-1011.

[22] Liu, D. and Gildea, D. (2008), Improved tree-to-string transducer for machine translation, *In Proceedings of the Third Workshop on Statistical Machine Translation*, pp. 62-69.

[23] Hopkins, M. and Kuhn, J. (2007). Machine translation as tree labeling, *In Proceedings of SSST, NAACL-HLT 2007/AMTA Workshop on Syntax and Structure in Statistical Translation*, pp. 41-48.

[24] G. Kikui, E. Sumita, T. Takezawa, and S. Yamamoto. (2003). Creating corpora for

speech-to-speech translation. *In Proceeding of EUROSPEECH-03*, pp. 381-384.

[25] Slav Petrov, Leon Barrett, Romain Thibaux and Dan Klein. (2006). Learning Accurate, Compact, and Interpretable Tree Annotation, In Proceeding in COLING-ACL 2006, pp. 433-440.

[26] Win Pa Pa, Ye Kyaw Thu, Andrew Finch, Eiichiro Sumita. (2015). Word Boundary Identification for Myanmar Text Using Conditional Random Fields. In Proceedings of the Ninth International Conference on Genetic and Evolutionary Computing 2015, pp. 447-456

[27] Philipp Koehn and Barry Haddow. (2009). 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.

[28] F. J. Och and H. Ney. (2000). Improved statistical alignment models. *In ACL00*, pp. 440–447.

[29] Christoph Tillmann. (2004). A unigram orientation model for statistical machine translation. *In Proceedings of HLT-NAACL 2004*: Short Papers, HLT-NAACL-Short '04, pp. 101–104.

[30] Andreas Stolcke. (2002). SRILM - An Extensible Language Modeling Toolkit. *In Proceedings of the International Conference on Spoken Language Processing, volume 2*, pp. 901–904

[31] Franz J. Och. (2003). Minimum error rate training for statistical machine translation. *In Proceedings of the 41st Meeting of the Association for Computational Linguistics (ACL 2003)*, pp. 160-167.

[32] K. Papineni, S. Roukos, T. Ward, and W.J. Zhu. (2001). Bleu: a Method for Automatic Evaluation of Machine Translation. *IBM Research Report rc22176 (w0109022)*, Thomas J. Watson Re- search Center, 2001.

[33] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. (2010). Automatic evaluation of translation quality for distant language pairs. *In Proceedings of EMNLP'10*, pp. 944–952.

[34] L. Thurstone. (1927). The method of paired comparisons for social values. Journal of Abnormal and Social Psychology, 21:384-400.

[35] LDC. 2005. Linguistic data annotation specification: Assessment of fluency and adequacy in translations. Revision 1.5.

[36] Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C. and Schroeder, J. (2007). (Meta-) Evaluation of Machine Translation, *in Proceedings of the Second Workshop on Statistical Machine Translation*, pp. 136–158.

[37] Koehn, P. (2004). Statistical Significance Tests For Machine Translation Evaluation, *in Proceedings of EMNLP 2004*, pp. 388-395.

[38] B. Efron and R. J. Tibshirani. (1993). An Introduction to the Bootstrap. Chapman & Hall, New York and London.

[39] M. G. Kendall. (1938). A new measure of rank corre- lation. Biometrika, 30(1/2): pp. 81–93.