Hybrid Statistical Machine Translation for English-Myanmar: UTYCC Submission to WAT-2021

Ye Kyaw Thu[1,2], Thazin Myint Oo[2], Hlaing Myat Nwe[1], Khaing Zar Mon[1],
Nang Aeindray Kyaw[1], Naing Linn Phyo[1], Nann Hwan Khun[1], Hnin Aye Thant[1]
[1]University of Technology (Yatanarpon Cyber City), Myanmar
[2]Language Understanding Lab., Myanmar
{yekyawthu,hlaingmyatnwe,khaingzarmon,nangaeindraykyaw}@utycc.edu.mm,
{nainglinphyo,nannhwankhun, hninayethant}@utyccc.edu.mm,
queenofthazin@gmail.com

Abstract

In this paper we describe our submissions to WAT-2021 (Nakazawa et al., 2021) for English-to-Myanmar language (Burmese) task. Our team, ID: "YCC-MT1", focused on bringing transliteration knowledge to the decoder without changing the model. We manually extracted the transliteration word/phrase pairs from the ALT corpus and applying XML markup feature of Moses decoder (i.e. `-xml-input exclusive`, `-xml-input inclusive`). We demonstrate that hybrid translation technique can significantly improve (*around 6 BLEU scores*) the baseline of three well-known "Phrase-based SMT", "Operation Sequence Model" and "Hierarchical Phrase-based SMT". Moreover, this simple hybrid method achieved the second highest results among the submitted MT systems for English-to-Myanmar WAT2021 translation share task according to BLEU (Papineni et al., 2002) and AMFM scores (Banchs et al., 2015).

1 Introduction

While both statistical machine translation (SMT) and neural machine translation (NMT) have proven successful for high resource language, it is still an open research question how to make it work well especially for the low resource and long distance reordering language pairs such as English and Burmese (Duh et al., 2020), (Kolachina et al., 2012), (Trieu et al., 2019), (Win Pa Pa et al., 2016). To the best of our knowledge there are only two publicly available English-Myanmar parallel corpora; ALT Corpus (Ding et al., 2020) and UCSY Corpus (Yi Mon Shwe Sin and Khin Mar Soe, 2019) for research purpose, and the size of the corpora are around 20K and 200K respectively. The parallel data for Myanmar-English machine translation share task at Workshop on Asian Translation (WAT) using combination of that two corpora and thus it is a good chance for the NLP researchers who are working on low resource machine translation. Motivated by this challenge, we represented the University of Technology, Yatanarpon Cyber City (UTYCC) and participated in the English-Myanmar (en-my) share task of WAT2021 (Nakazawa et al., 2021).

In this paper, we propose one hybrid system based on plugging XML markup translation knowledge to the SMT decoder. The translation rules for transliteration and borrowed words, and direct usage of English words in the target language are constructed by using a parallel word dictionary. The English-Myanmar transliteration dictionary was built by manual extracting parallel words/phrases from the whole ALT corpus. This simple hybrid method outperformed the three baselines and achieved the second highest results among the submitted MT systems for English-to-Myanmar WAT2021 translation share task according to BLEU (Papineni et al., 2002) and AMFM scores (Banchs et al., 2015).

The remainder of this paper is organized as follows. In Section 2, we introduce the data preprocessing, including word segmentation and cleaning steps. In Section 3, we describe the details of our three SMT systems. The machine translation evaluation metrics are presented in Section 4. The manual extraction process of transliteration word/phrase pairs from the ALT English-Myanmar parallel data is described in Section 5. Then, the SMT decoding with XML markup technique is described in Section 6. In Section 7, we present hybrid translation results achieved by all our systems. Section 8 concludes this paper.

2 Data preprocessing

2.1 Preprocessing for English and Myanmar

We tokenized and escaping English data respectively with the tokenizer and escaping perl script (`escape-special-chars.perl`) of Moses (Koehn et al., 2007). For Myanmar, although provided training data of ALT was already segmented, word segmentation was not provided for the UCSY corpus. And thus, we did syllable segmentation by using `sylbreak.pl` (Ye Kyaw Thu, 2017).

2.2 Parallel Data Statistic

The corpus for the English-Myanmar share task contained two separate corpora and they are UCSY corpus and ALT corpus. The domain of the UCSY corpus is general and the

Table 1: Statistics of our preprocessed parallel data

Data Type	# of Sentences	# of Myanmar Syllables	# of English Words
TRAIN (UCSY)	238,014	6,285,996	3,357,260
TRAIN (ALT)	18,088	1,038,640	413,000
DEV	1,000	57,709	27,318
TEST	1,018	58,895	27,929

original English sentences of the ALT corpus was extracted from the Wikinews (Ye Kyaw Thu et al., 2016). The size of the UCSY parallel corpus is about 238K sentence pairs and it is also a part of the training data of the WAT2021 share task. The size of the English-Myanmar ALT parallel corpus is about 20K sentence pairs and splitted into 18,088 sentences for training, 1,000 sentences for development and 1,018 sentences for test data respectively (see Table 1). While the number of development and test set sentences are the same, we implemented phrase-based, operating sequence model and hiero systems with the UCSY training corpus only and the combination of the UCSY and the ALT training sets. The statistics of our preprocessed parallel data are shown in Table1.

3 SMT Systems

In this section, we describe the methodology used in the machine translation experiments for this share task.

3.1 Phrase-based Statistical Machine Translation

A PBSMT translation model is based on phrasal units (Koehn et al., 2003). 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 corpus and a basic reordering model, and an algorithm to extract the phrases to build a phrase-table (Specia, 2011). The phrase translation model is based on noisy channel model. To find best translation $\hat{e}$ that maximizes the translation probability $\mathbf{P}(f)$ given the source sentences; mathematically. Here, the source language is French and the target language is an English. The translation of a French sentence into an English sentence is modeled as equation 1.

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

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

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

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

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

3.2 Operation Sequence Model

The operation sequence model which combines the benefits of two state-of-the-art SMT frameworks named n-gram-based SMT and phrase-based SMT. This model simultaneously generate source and target units and does not have spurious ambiguity that is based on minimal translation units (Durrani et al., 2011) (Durrani et al., 2015). It is a bilingual language model that also integrates reordering information. OSM motivates better reordering mechanism that uniformly handles local and non-local reordering and strong coupling of lexical generation and reordering. It means that OSM can handle both short and long distance reordering. The operation types are such as generate, insert gap, jump back and jump forward which perform the actual reordering.

3.3 Hierarchical Phrase-based Statistical Machine Translation

The hierarchical phrase-based SMT approach is a model based on synchronous context-free grammar (Specia, 2011). The model is 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 language pairs that require long-distance re-ordering during the translation process (Braune et al., 2012).

3.4 Moses SMT System

We used the PBSMT, HPBSMT and OSM system provided by the Moses toolkit (Koehn et al., 2007) for training the PBSMT, HPBSMT and OSM statistical machine translation systems. The word segmented source language was aligned with the word segmented target language using GIZA++ (Och and Ney, 2000). The alignment was symmetrized by grow-diag-final and heuristic (Koehn et al., 2003). The lexicalized reordering model was trained with the msd-bidirectional-fe option (Tillmann, 2004). We use KenLM (Heafield, 2011) for training the 5-gram language model with modified Kneser-Ney discounting (Chen and Goodman, 1996). Minimum error rate training (MERT) (Och, 2003) was used to tune the decoder parameters and the decoding was done using the Moses decoder (version 2.1.1). We used default settings of Moses for all experiments.

4 Evaluation

Our systems are evaluated on the ALT test set and we used the different evaluation metrics such as Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002), Rank-based Intuitive Bilingual Evaluation Score (RIBES) (Isozaki et al., 2010), and Adequacy-Fluency Metrics (AMFM) (Banchs et al.,

2015). For the official evaluation of English-to-Myanmar share task, we uploaded our hypothesis files to the WAT2021 evaluation server and sub-syllable (almost same with sylbreak toolkit's syllable units) segmentation was used for Myanmar language. We submitted "hybrid PBSMT with XML markup (inclusive)" and "hybrid OSM with XML markup (inclusive)" systems training only with UCSY corpus for human evaluation.

5 Manual Extraction of Parallel Transliteration Words

When we studied on the Myanmar language corpus provided by the WAT2021, we found that many sentences are very long, containing spelling errors, unnaturalness of translation (i.e. translation from English to Myanmar) and many transliteration words (i.e. word-by-word, phrase-by-phrase, compound word transliteration). Moreover, the ALT corpus was extracted from the English Wikinews (Ye Kyaw Thu et al., 2016) and it contains many named entity words such as person names, organizations, locations. See three example English-Myanmar parallel sentences of the ALT test data that contained several transliteration words. This paper tackling the problem of machine translation on transliteration word/phrase pairs by hybrid translation approach.

[ဆစ် ဒ နီ]¹ က [ရ နဲ့ ဝစ်ခ်]² မြင်း ပြိုင် ကွင်း မှ မျိုး သ နဲ့ ပြိုင် မြင်း ရှစ် ကောင် ဟာ မြင်း တုတ် ကွေး ရော ဂါ ကူး စက် ခံ ခဲ့ ရ တယ် ဆို တာ အ တည် ပြု ခဲ့ ပါ တယ် ။ (It has been confirmed that eight thoroughbred race horses at [Randwick]² Racecourse in [Sydney]¹ have been infected with equine influenza .)

ဒီ စ နေ တ နင်္ဂ နွေ မှာ [အန် အက်စ် ဒ ဗ လျူ]¹ နှ င့် [ကွင်း စ် လန်း ဒ်]² မှ လွဲ ၍ [ဩ စ တြေး လျ]³ တိုက် ပြည် နယ် များ အား လုံး တွင် မြင်း ပွဲ ပြန် စ ဖို့ မျှော် လ င့် ပါ တယ် ။ (Racing is expected to resume in all [Australian]³ states except [NSW]¹ and [Queensland]² on the weekend .)

[ဘ ရစ် တ နီ စ ပီး ယား]¹ သူ မ ၏ နေ အိမ် သို့ အ ပြန် ကို နောက် ယောင် ခံ လိုက် ပြီး နောက် [ကယ် လီ ဖိုး နီး ယား]² ပြည် နယ် [မစ် ရှင်း]³ တောင် ကုန် များ အ တွင်း ၊ အ တား အ ဆီး မဲ့ စွာ မောင်း နှင် ခြင်း အ တွက် ဗုဒ္ဓ ဟူး နေ့ တွင် သ တင်း ထောက် အ ဖွဲ့ ဝင် အ မျိုး သား လေး ယောက် အ ဖမ်း ခံ ခဲ့ ရ ပြီး ဒဏ် ငွေ ပေး ဆောင် ခဲ့ ရ သည် ။ (Four male members of the paparazzi were arrested and charged on Wednesday with reckless driving in [Mission]³1 Hills , [California]² after following [Britney Spears]¹ back to her mansion .)

We manually extracted English-Myanmar transliteration word and phrase pairs from the whole ALT corpus and prepared 14,225 **unique word dictionary**[1]. The main categories are Country/City Names, Demonyms, Personal Names, Month Names, General Nouns, Organization Names, Abbreviations, Units and English-to-English Translation Words (see Table 2).

6 Hybrid Translation

Generally, hybrid translation integrates the strengths of rationalism method and empiricist method. (Hunsicker et al., 2012) described how machine learning approaches can be used to improve the phrase substitution component of a hybrid machine translation system. Essential of hybrid translation is to integrate the

---

[1] https://github.com/ye-kyaw-thu/MTRSS/tree/master/WAT2021/en-my_transliteration-dict

Table 2: Some example of manually extracted transliteration word/phrase pairs.

Country/City Names and Demonyms	
Italy	အီတလီ
Portugal	ပေါ်တူဂီ
Paris	ပြင်သစ်
Shanghai	ရှန်ဟိုင်း
Australian	သြစတြေးလျ လူမျိုး
Personal Names	
David Bortolussi	ဒေးဗစ် ဘော်တိုလပ်စီ
coach William McKenny	ကော့ချ် ဝီလီယမ် အမ်စီကန်နီ
Dr. Michel Pellerin	ဒေါက်တာ မီရှဲလ် ပဲလာရီ
Liu Jianchao	လျူ ကျန်းချောင်
manager Phil Garner	မန်နေဂျာ ဖီး ဂါနာ
Month Names	
January	ဇန်နဝါရီ လ
March	မတ် လ
May	မေ လ
September	စက်တင်ဘာ
October	အောက်တိုဘာ
General Nouns	
penalties	ပယ်နယ်လ်တီ
theory	သီအိုရီ
bowling	ဘိုးလင်း
the Yankees	ရန်ကီး
baseball	ဘေ့စ်ဘော
Organization Names	
Liberal Democrats	လစ်ဘရယ် ဒီမိုကရက်
Scottish Premier League (SPL)	စကော့တလန် ပရီမီယာ လိဂ် (အက်စ်ပီအယ်လ်)
Walt Disney World's Wide World of Sports	ဝေါ့ ဒစ္စနေး ဝေါလ် ၏ ဝိုက် ဝါ စ ပေါ့
Somali Defence Ministry	ဆိုမာလီ ကာကွယ်ရေး ဝန်ကြီး ဌာန
Iranian press-agency IRNA	အီရန် စာနယ်ဇင်း -အေဂျင်စီ အာနာ
Abbreviations and Units	
Intel x86	အင်တဲလ် အိတ်စ်၈၆
NFL	အန်အက်ဖ်အယ်
A.Q.U.S.A	အေ.ကျူ.ယူ.အက်.အေ
AC-130	အေစီ-၁၃၀
km	ကီလိုမီတာ
English-to-English Words	
Big C	Big C
iTV	iTV
F-16	F - 16
Khlong Toei	Khlong Toei
Na Ranong	Na Ranong

86

core of MT engines. Multiple- engine HMT integrates all available MT methods, applying to their benefits, in order to improve qualities of output (Xuan et al., 2012). The popular combinations comprise "rule-based machine translation vs the SMT" and multiple combinations of machine translation engines, for example "SMT vs neural machine translation". Our work in this paper focuses on hybrid machine translation of SMT engine and XML tags inserting (i.e. applying rules) into transliteration words of each source sentence. We used the Moses SMT toolkit and it also supports `-xml-input` flag to activate XML tags inserting feature with one of the five options; `exclusive`, `inclusive`, `constraint`, `ignore` and `pass-through`. Refer manual page of the Moses toolkit [2] for detail explanation. Although we studied all options, we will present the two options that work well for English-Myanmar hybrid translation.

The Moses decoder has an XML markup scheme that allows the specification of translations for parts of the sentence. In its simplest form, we can guide the decoder what to use to translate certain transliteration words or phrases in the source sentence. We wrote a perl script for XML Markup inserting into the source English sentences based on the manually extracted transliteration dictionary. As shown in follows, the XML Markup scheme for HPBSMT is different with PBSMT and OSM. This is because the syntactic annotation of the HPBSMT system also used XML Markup. And thus, we used `--xml-brackets "{{ }}"` option when decoding hybrid HPB-SMT system.

XML Markup Scheme for PBSMT and OSM:
Tanks of <np translation="အောက် စီ ဂျင်" prob="0.8">oxygen</np> , <np translation="ဟီ လီ ရမ်" prob="0.8">helium</np> and <np translation="အက် ဆီ တ လင်း" prob="0.8">acetylene</np> began to explode after a connector used to join <np translation="အက် ဆီ တ လင်း" prob="0.8">acetylene</np> tanks during the filling process malfunctioned .

Decoding with XML Markup Scheme for PBSMT and OSM:
```
$moses -xml-input exclusive -i ./test.xml.en -f ../evaluation/test.filtered.ini.1
> en-my.xml.hyp1
```

XML Markup Scheme for HPBSMT:
Tanks of {{np translation="အောက် စီ ဂျင်" prob="0.8"}}oxygen{{/np}} , {{np translation="ဟီ လီ ရမ်" prob="0.8"}}helium{{/np}} and {{np translation="အက် ဆီ တ လင်း" prob="0.8"}}acetylene{{/np}} began to explode after a connector used to join {{np translation="အက် ဆီ တ လင်း" prob="0.8"}}acetylene{{/np}} tanks during the filling process malfunctioned .

Decoding with XML Markup Scheme for HPBSMT:
```
$moses_chart -xml-input exclusive --xml-brackets "{{ }}" -i ./test.xml.en -f
../evaluation/test.filtered.ini.1 > en-my.xml.hyp1
```

7 Results

Our systems are evaluated on the ALT test set and the results are shown in Table 3. Our observations from the results are as follows:

1. Hybrid translation of SMT with XML Markup scheme showed significant improvement for all three SMT approaches; PBSMT, OSM and HPBSMT.

2. Generally, `-xml-input exclusive` option gives a slightly higher scores than `-xml-input inclusive`.

3. HPBSMT achieved the highest scores especially for training without ALT corpus (i.e. we can assume working well for Out-of-Vocabulary case).

4. The baseline translation performance score difference between training with or without ALT corpus is about 5.0 BLEU score.

8 Conclusion

We presented in this paper the UTYCC's participation in the WAT-2021 shared translation task. Our hybrid SMT submission to the task performed the second in English-to-Myanmar translation direction according to several evaluation scores including the de facto BLEU. Our results also confirmed the XML markup technique for transliteration words dramatically increase the translation performance up

---

Table 3: BLEU, RIBES and AMFM scores for English-to-Myanmar translation (Bold number indicate the highest score for each scoring method)

Experiments	Only UCSY Training Data			UCSY+ALT Training Data		
	BLEU	RIBES	AMFM	BLEU	RIBES	AMFM
Baseline: PBSMT	15.01	0.519451	0.550400	20.80	0.542406	0.617900
Hybrid: xml-exclusive	20.80	0.551514	0.653850	24.54	0.563854	**0.690020**
Hybrid: xml-inclusive	**20.88**	**0.553319**	**0.655310**	**25.11**	**0.567187**	0.689400
Baseline: OSM	15.05	0.528968	0.557240	20.33	0.550329	0.622350
Hybrid: xml-exclusive	19.94	0.540820	0.651070	**23.82**	0.554226	0.691450
Hybrid: xml-inclusive	**20.13**	**0.545962**	**0.654820**	23.73	**0.556381**	**0.691910**
Baseline: Hiero	14.83	0.555290	0.545900	20.29	0.587136	0.612400
Hybrid: xml-exclusive	21.02	0.588198	0.653840	25.48	0.60733	0.684110
Hybrid: xml-inclusive	21.02	0.588198	0.653840	25.48	0.607339	0.684110

to 6 BLEU scores. Moreover, our results highlighted that hybrid statistical machine translation is easy to implement and we need to explore more for low-resource distant language pairs such as English-Myanmar translation.

References

Rafael E. Banchs, Luis F. D'Haro, and Haizhou Li. 2015. Adequacy–fluency metrics: Evaluating mt in the continuous space model framework. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):472–482.

Fabienne Braune, Anita Gojun, and Alexander Fraser. 2012. Long-distance reordering during search for hierarchical phrase-based SMT. In *Proceedings of the 16th Annual conference of the European Association for Machine Translation*, pages 177–184, Trento, Italy. European Association for Machine Translation.

Stanley F. Chen and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, ACL '96, page 310–318, USA. Association for Computational Linguistics.

Chenchen Ding, Sann Su Su Yee, Win Pa Pa, Khin Mar Soe, Masao Utiyama, and Eiichiro Sumita. 2020. A Burmese (Myanmar) treebank: Guildline and analysis. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(3):40.

Kevin Duh, Paul McNamee, Matt Post, and Brian Thompson. 2020. Benchmarking neural and statistical machine translation on low-resource African languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2667–2675, Marseille, France. European Language Resources Association.

Nadir Durrani, Helmut Schmid, and Alexander Fraser. 2011. A joint sequence translation model with integrated reordering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1045–1054, Portland, Oregon, USA. Association for Computational Linguistics.

Nadir Durrani, Helmut Schmid, Alexander Fraser, Philipp Koehn, and Hinrich Schütze. 2015. The operation sequence Model—Combining n-gram-based and phrase-based statistical machine translation. *Computational Linguistics*, 41(2):157–186.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Sabine Hunsicker, Yu Chen, and Christian Federmann. 2012. Machine learning for hybrid ma-

chine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 312–316, Montréal, Canada. Association for Computational Linguistics.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952, Cambridge, MA. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

Prasanth Kolachina, Nicola Cancedda, Marc Dymetman, and Sriram Venkatapathy. 2012. Prediction of learning curves in machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22–30, Jeju Island, Korea. Association for Computational Linguistics.

Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, and Sadao Oda, Yusuke Kurohashi. 2021. Overview of the 8th workshop on Asian translation. In *Proceedings of the 8th Workshop on Asian Translation*, Bangkok, Thailand. Association for Computational Linguistics.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 440–447, Hong Kong. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Lucia Specia. 2011. Tutorial, fundamental and new approaches to statistical machine translation. *International Conference Recent Advances in Natural Language Processing*.

Christoph Tillmann. 2004. A unigram orientation model for statistical machine translation. In *Proceedings of HLT-NAACL 2004: Short Papers*, pages 101–104, Boston, Massachusetts, USA. Association for Computational Linguistics.

Hai-Long Trieu, Duc-Vu Tran, Ashwin Ittoo, and Le-Minh Nguyen. 2019. Leveraging additional resources for improving statistical machine translation on asian low-resource languages. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 18(3):32:1–32:22.

Win Pa Pa, Ye Kyaw Thu, Andrew Finch, and Eiichiro Sumita. 2016. A study of statistical machine translation methods for under resourced languages. *Procedia Computer Science*, 81:250–257. SLTU-2016 5th Workshop on Spoken Language Technologies for Under-resourced languages 09-12 May 2016 Yogyakarta, Indonesia.

H.W. Xuan, W. Li, and G.Y. Tang. 2012. An advanced review of hybrid machine translation (hmt). *Procedia Engineering*, 29:3017–3022. 2012 International Workshop on Information and Electronics Engineering.

Ye Kyaw Thu. 2017. sylbreak toolkit for burmese (myanmar language). Accessed: 2021-03-06.

Ye Kyaw Thu, Win Pa Pa, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2016. Introducing the asian language treebank (alt). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Yi Mon Shwe Sin and Khin Mar Soe. 2019. Attention-based syllable level neural machine translation system for myanmar to english language pair. *International Journal on Natural Language Computing*, 8(2):01–11.