Statistical Machine Translation between Myanmar Written Text (MWT) and Myanmar SignWriting (MSW)

Second Seminar

Supervisor: Dr. Nandar Win Min
Co-supervisor: Dr. Ye Kyaw Thu, Daw Swe Zin Moe
Presented by: Hnin Wai Wai Hlaing (ME-IST 23)
NLP Research Group

(9th February 2018)

Outlines

- Abstract
- Objectives
- Introduction
- Contribution
- Proposed System Design
- Myanmar sign language
- SignWriting (SW)
- Experiment and Evaluation
- Conclusion
- References

Abstract

- In the field of machine translation, significant progress has been made by using statistical methods.
- The proposed system suggests a **statistical machine translation system between Myanmar Written Text and Myanmar SignWriting.**
- It takes Myanmar Written Text as input and the output is in the form of Myanmar SignWriting.
- There is no Myanmar Written Text and Myanmar SignWriting parallel data yet, and thus it is needed to prepare.
- It solves difficulties for deaf people to learn the basic concept of daily life, especially in **emergency case**.

Keywords: *Myanmar SignWriting, Statistical Machine Translation, Natural Language Processing.*

Objectives

- To learn Machine Translation between Myanmar Written Text and Myanmar SignWriting
- To develop Myanmar Written Text and Myanmar SignWriting parallel corpus
- To measure Machine Translation performance using Statistical Machine Translation (SMT) approaches
- To fulfill the communication requirements between deaf people and hearing people

Introduction

- Sign language is the natural language of the Deaf and thus they have some problems in communicating and knowledge sharing with hearing people.
- Myanmar sign language is used as a primary means of communication for Myanmar deaf people, about 1.3% of population in Myanmar.
- As they are limited resources of information written in their language, Myanmar SignWriting translation system is very important for the Deaf.
- This proposed system focus on machine translation between Myanmar written text and Myanmar SignWriting (MSW).

Contribution

- It is the first evaluation of the Statistical Machine Translation approaches between Myanmar Written Text and Myanmar SignWriting.
- It is to build Myanmar Written Text and Myanmar SignWriting parallel corpus and this will be useful for further researches.
- It will investigate on statistical machine translation (SMT) performance between Myanmar written text and Myanmar SignWriting.

System Design

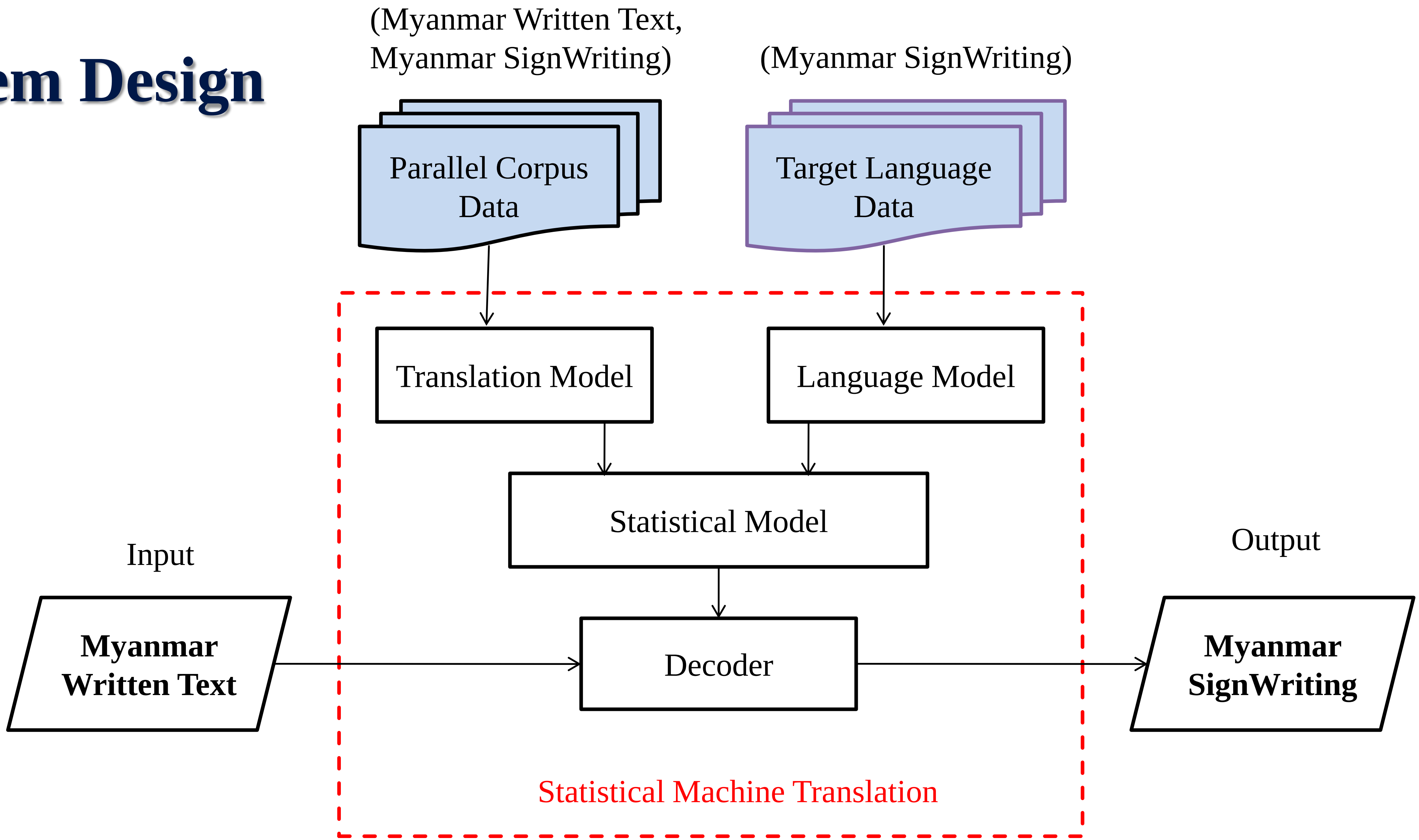


Figure 1: Flow Diagram of the proposed system

Sign Language (SL)

- SL is the native language of the Deaf community.
- It is a **vision-based language** as Deaf can see.
- They can express their needs and the formation of concepts by combining hand shapes, orientation and movement of the hands, arms or body, and facial expressions.
- SL consists of **Manual Features (MFs),and Non-Manual Features (NMFs).**

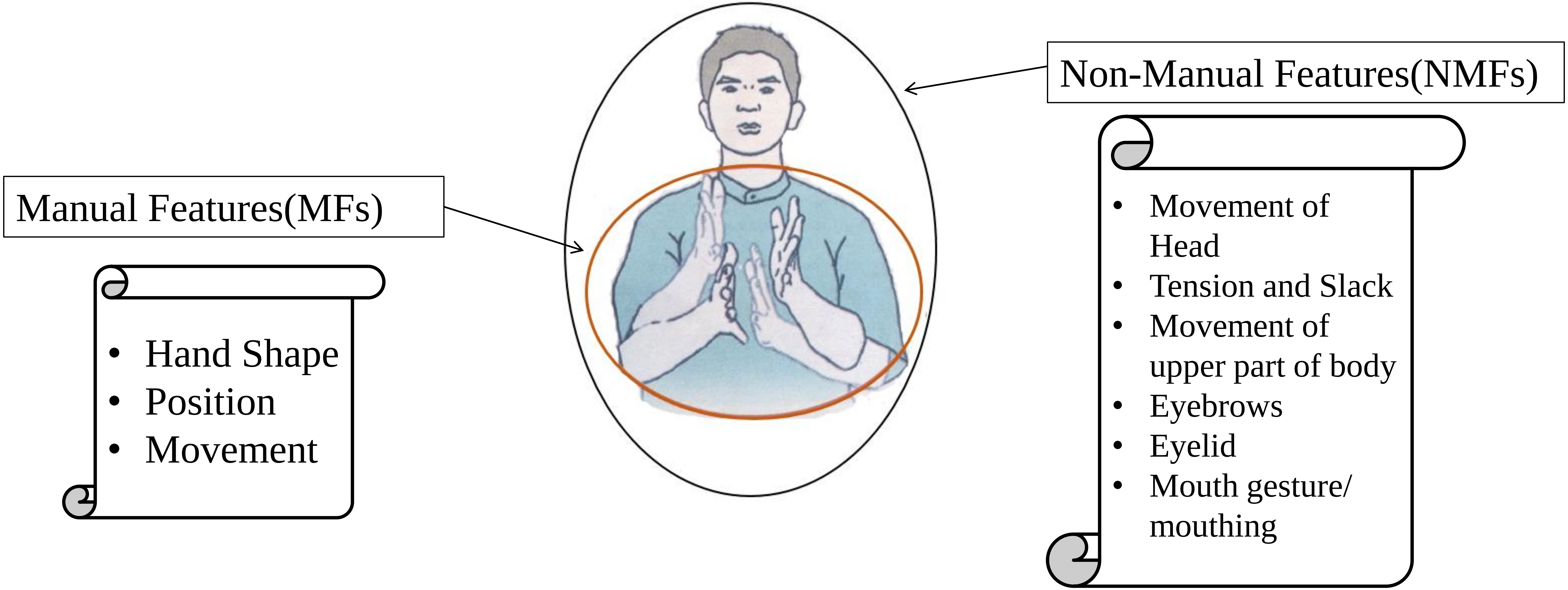


Figure 2: Structure of Sign Language

Myanmar Sign Language (MSL)

- Each country has its own, native sign language according to their culture.
- There are four schools for the Deaf in Myanmar.
- MSL is a primary communication for Myanmar Deaf community.
- MSL has its own grammar structure which is very difference with Myanmar written text.
- A number of written systems for representing sign languages have been developed, and defined with SignWriting Alphabets for each country.
- Myanmar SignWriting for the Deaf is needed to define for each sign.

SignWriting (SW)

- SW is a writing system that is a sequence of symbols for deaf sign language.
- Deaf represents two perspective: **singer’s perspective** and **observer’s perspective.**
- SignWriting is based on how you see your own hands when you sign—the **signer’s perspective.**
- SW is written horizontally **(left to right)** and the **right hand is dominant**.
- SW symbols can be rotated in 8 directions and placed anywhere in the writing area.
- **International Sign Writing Alphabet (ISWA) 2010** defines **7 categories, 30 groups** of symbols to form **652 base symbols** and **35,023 final symbols**.

Category 1: Hands	Group 01: Index Group 02: Index Middle Group 03: Index Middle Thumb Group 04: Four Fingers Group 05: Five Fingers	Group 06: Baby Finger Group 07: Ring Finger Group 08: Middle Finger Group 09: Index Thumb Group 10: Thumb
Category 2:Movement	Group 11: Contact Group 12: Finger Movement Group 13: Straight Wall Plane Group 14: Straight Diagonal Plane Group 15: Straight Floor Plane	Group 16: Curves Parallel Wall Plane Group 17: Curves Hit Wall Plane Group 18: Curves Hit Floor Plane Group 19: Curves Parallel Floor Plane Group 20: Circles
Category 4 :Dynamics & Timing	Group 21: Dynamics & Timing	
Category 5 :Body	Group 27: Trunk Group 28: Limbs	
Category 6: Detailed Location	Group 29: Detailed Location	
Category 7: Punctuation	Group 30: Punctuation	

SignWriting (Cont'd)

Front View

Top View

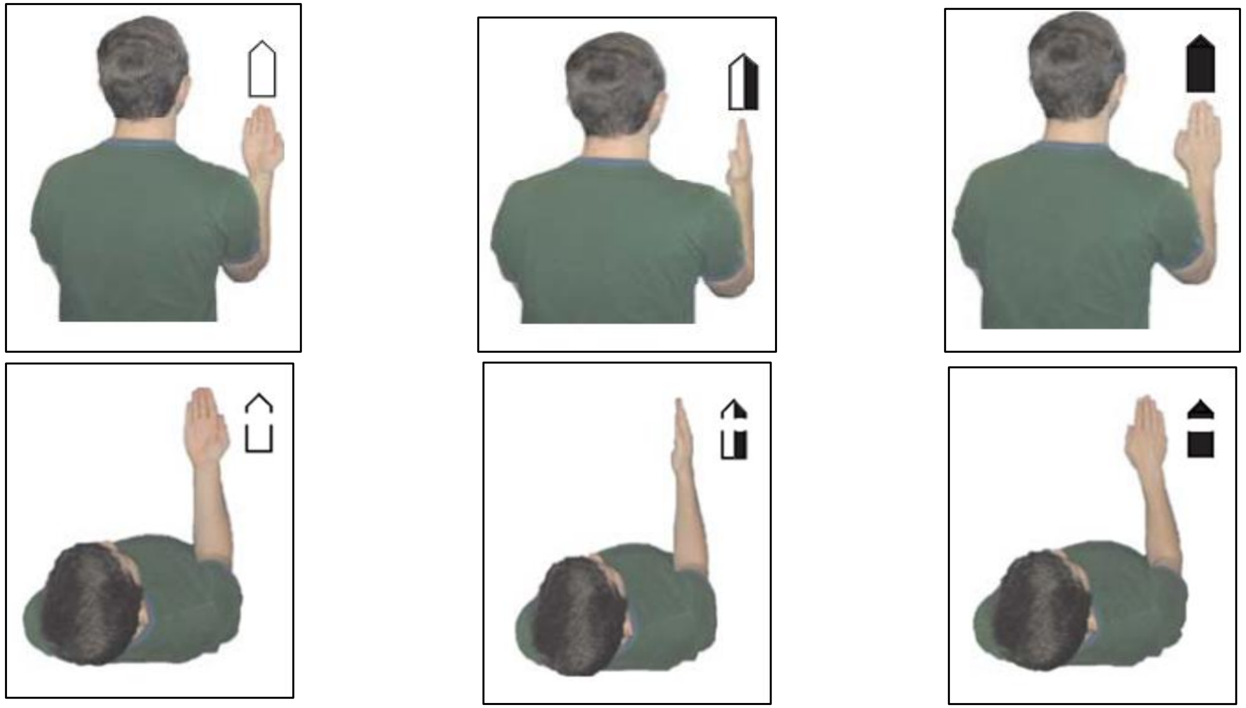


Figure 3: Example of SignWriting HAND-FLAT hand shapes

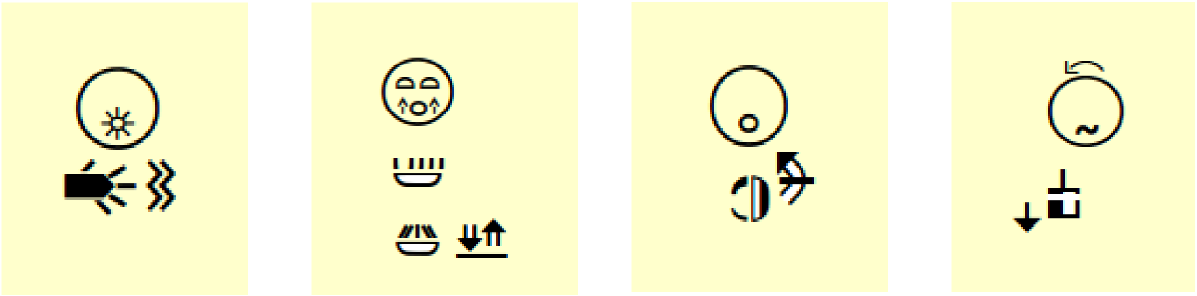
Transcription of Myanmar SignWriting

Myanmar Text : ကျိုချက် ထား သော ရေ ကို သောက်ပါ ။

Sign Language : ရေ ကျို သောက် ရ ။

Sign Video : idx20-128.mp4

SignWriting



Building Myanmar Written Text and Myanmar SignWriting parallel corpus

- The corpus contains Myanmar written text and the transcribed Myanmar SignWriting.
- There are many **challenges** in building the parallel corpus for SignWriting
  - Tokenization is manual because Myanmar written text does not contain tokenized characters (space, full stop, comma, etc.)
  - There is lack of Myanmar sign languages data collected.
  - SignWriting symbols need to clearly define for each Sign.
  - Myanmar Deaf and Myanmar sign language signers do not widely use SignWriting.

Building Myanmar Written Text and Myanmar SignWriting parallel corpus(Cont'd)

- Building Myanmar SignWriting parallel data contains **two parts.**
  - Data Collection and
  - Data Preparation.
- **Data Collection :**
  - The spoken style sentences and the written style sentences are manually selected from pamphlets and books for emergency situations.
  - SL trainer discussed with native signers and deaf persons to ensure the meaning of the original Myanmar written sentence using MSL.
  - After making discussion, video data are collected for each Myanmar sentence.
- **Data Preparation:**
  - Each sign can be defined by looking the recorded video in detail.(for both Manual and Non-manual signs)
  - After that, sign symbols are placed on the plane of SignMaker to form the shape and motion of Signs. (See Figure 4)

Building Myanmar Written Text and Myanmar SignWriting parallel corpus(Cont'd)

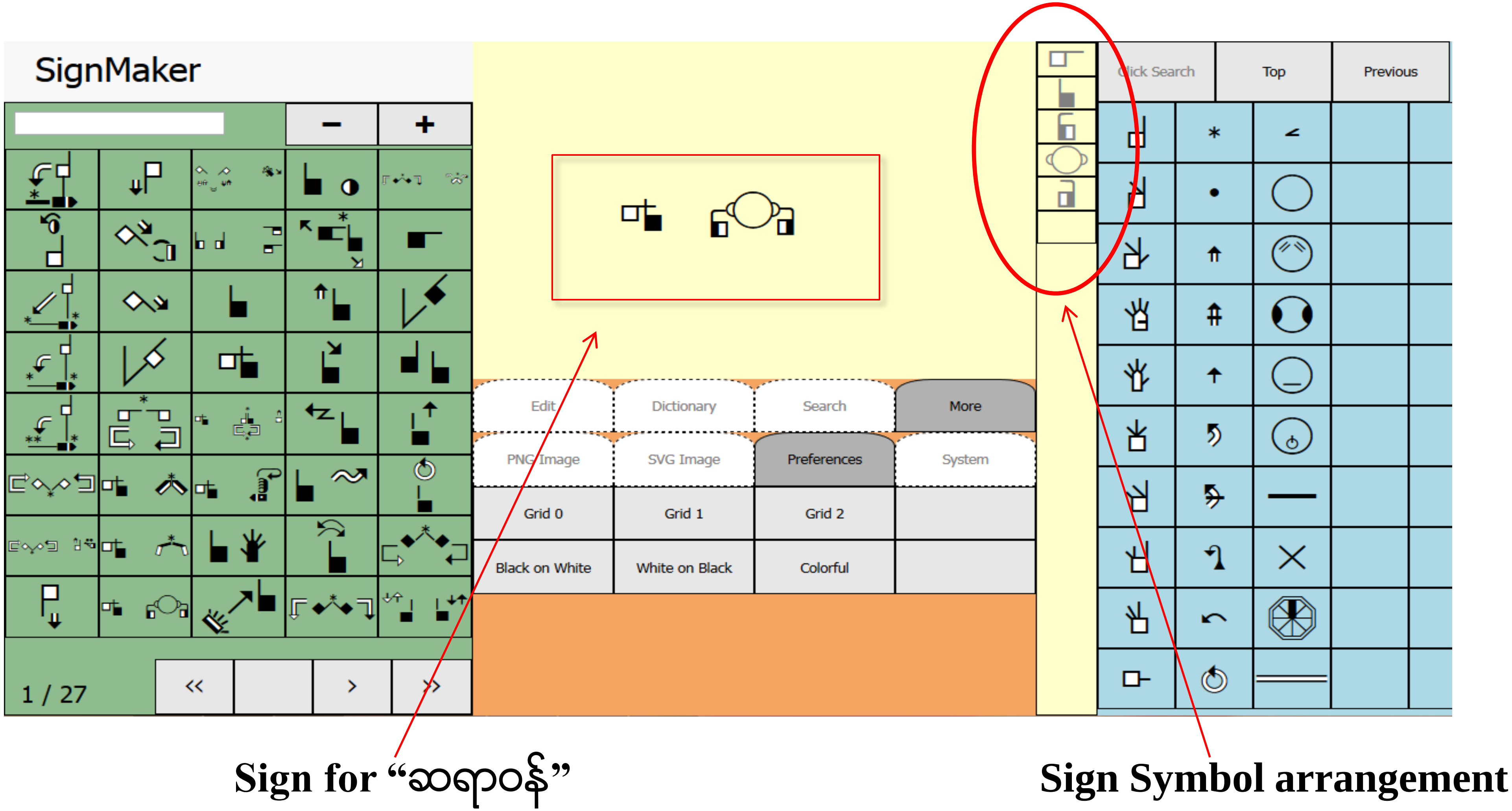


Figure 4: Example of SignWriting symbol preparation

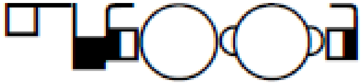
Building Myanmar Written Text and Myanmar SignWriting parallel corpus(Cont'd)

- By seeing the shape and movement of sign symbols in SignMaker, Unicode representation for each sign are arranged as a sentence.

- Example: Unicode representation of "ဆရာဝန်"

  - \U1D800\U1DAAA\U1D800\U1DA9C\U1D80A\U1DA9B\U1DAA8\U1D9FF\U1DA30\U1D80A\U1DA9B

- The final step is to transform the above Unicode sequence into the Sign symbol sequence.
- This step is needed because there is no input interface represented with SignWriting symbols.

-  shows the SignWriting representation of "ဆရာဝန်"

Building Myanmar Written Text and Myanmar SignWriting parallel corpus(Cont’d)

မီး ငြှိမ်း ။

မီးချိတ် ။

မီးကပ် ။

ရေ ပုံး ။

သဲ အိတ် ။

မီးသတ်ဆေးဘူး ။

မီးသတ်ရေကန် ။

မီးလောင် လွယ် သော ပစ္စည်း များ ။

လောင်စာဆီ ။

အမှိုက် ။

Myanmar Written Text **Myanmar SignWriting**

Statistical Machine Translation (SMT)

- SMT is a machine translation paradigm where translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text copra.
- There are many kinds of statistical machine translation approaches.
  - Word-Based Statistical Machine Translation
  - **Phrase-Based Statistical Machine Translation,**
  - Syntax-Based Statistical Machine Translation,
  - **Hierarchical Phrase-Based Statistical Machine Translation,**
  - **Operation Sequence Model.**

SMT (Cont'd)

- SMT approach can be described as modeling the probability distribution P(e|f), where e is a string in the source language and f is a string in the target language.
- Using Bayes' Rule, this can be rewritten as;

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

$$\hat{e} = \mathrm{argmax}_e\, P(e|f) = \mathrm{argmax}_e\, P(f|e)\, P(e)$$

  - P(f|e) : Translation Model
  - P(e) : Language Model
  - argmax_e : Decoder

SMT (Cont'd)

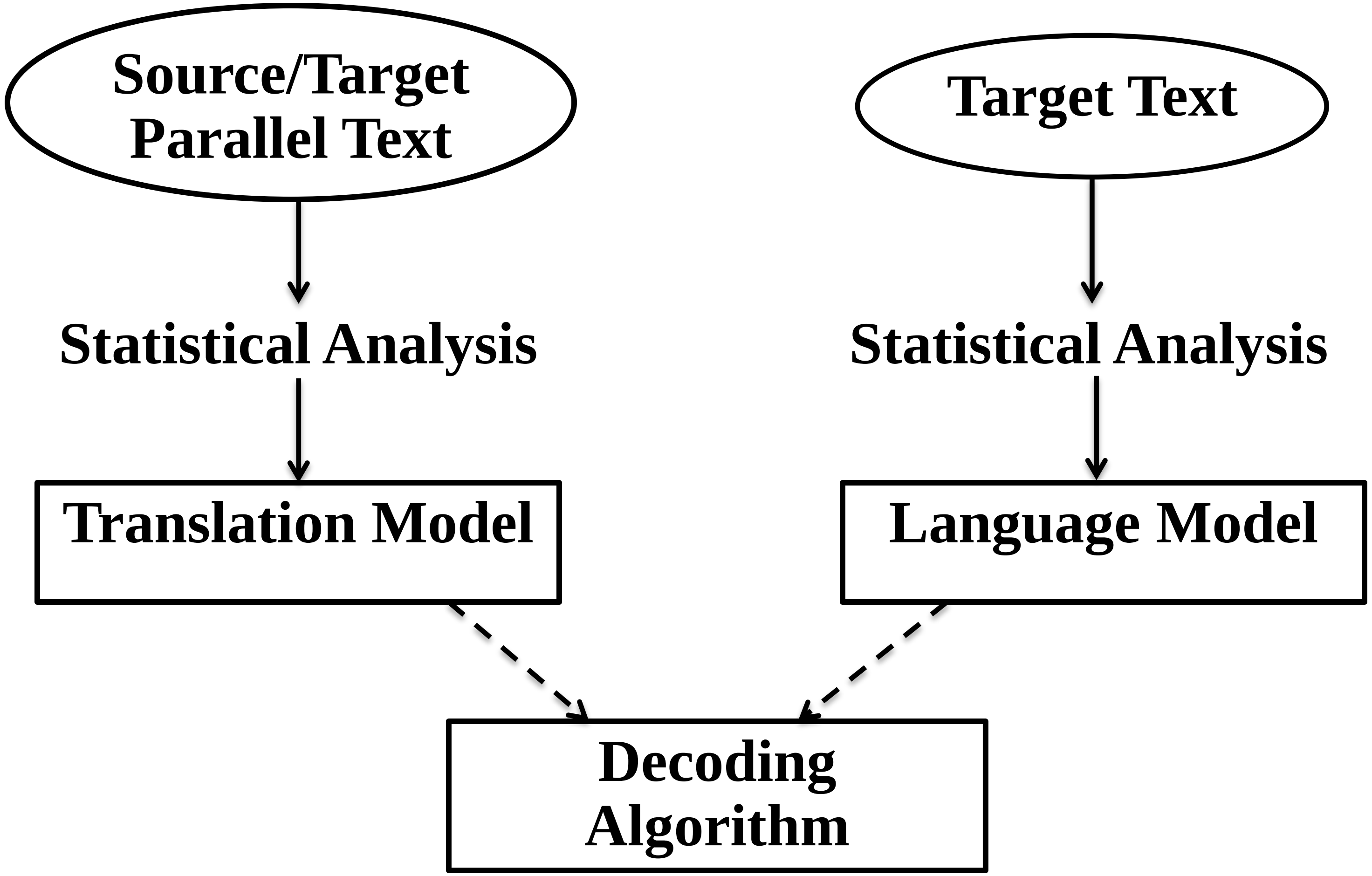


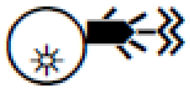
Figure 5: Components of SMT

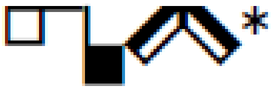
Phrase-Based Statistical Machine Translation (PBSMT)

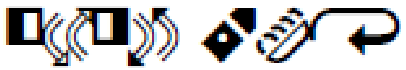
- PBSMT translate **phrases** as atomic units.
  - Phrase is a continuous sequence of words.
  - It is **not necessarily a linguistic phrase**.
- It is better translation performance than word-based .
- It consists of
  - **Phrase-pair probabilities** extracted from corpus,
  - **Reordering model**, and
  - An algorithm to extract the phrases to **build a phrase-table**.

PBSMT (Cont’d)

- First stage in training a phrase-based (PB) model is extraction of PB lexicon
- A PB lexicon pairs strings in one language with string in another language, e.g.,

ရေ ⟷ 

ဆေးရုံ ⟷ 

အရေးပေါ်ကား ⟷ 

… …

PBSMT (Cont'd)

Finding Alignment Matrices

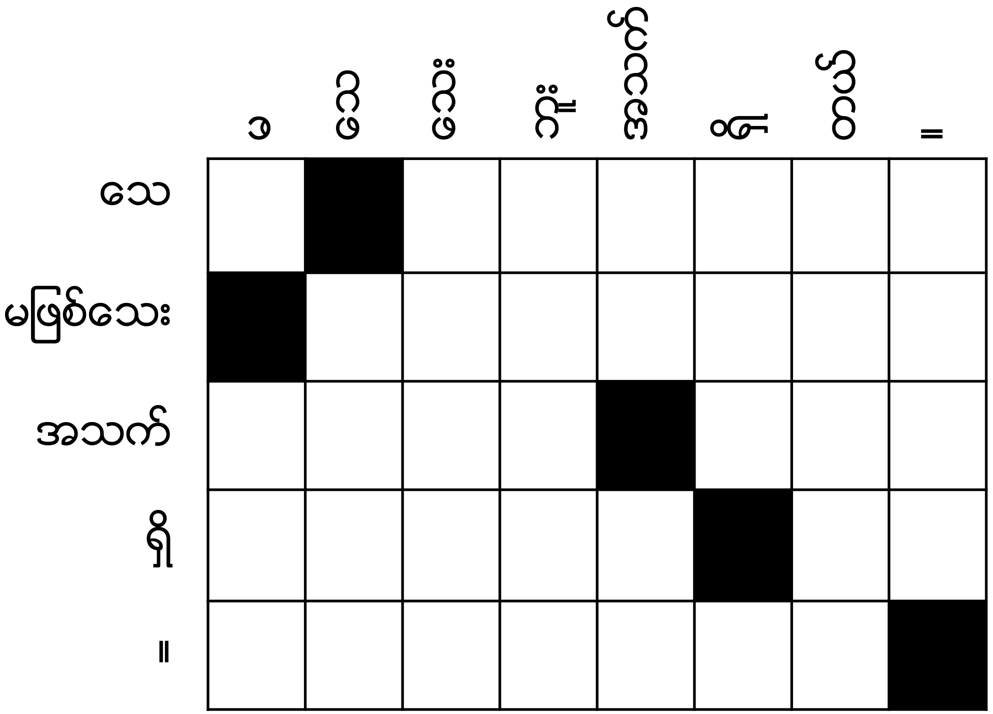
- Step 1: train IBM Model 2 for $P(f|e)$, and find the most likely alignment for each (e, f) pair
- Step 2: train IBM Model 2 for $P(e|f)$, and find the most likely alignment for each (f, e) pair
- Given the two alignments, take the intersection of the two as a starting point.

Myanmar Text (e) မ သေ သေး ဘူး အသက် ရှိ တယ် ။

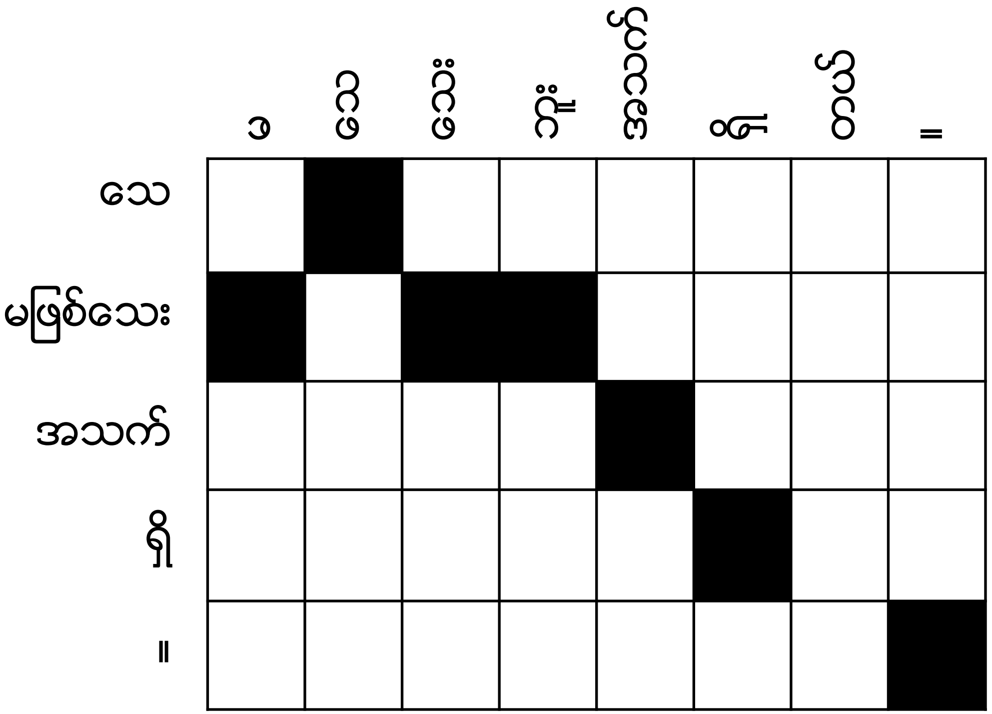
Sign Text (f) သေ မဖြစ်သေး အသက် ရှိ ။

PBSMT (Cont'd)

Finding Alignment Matrices



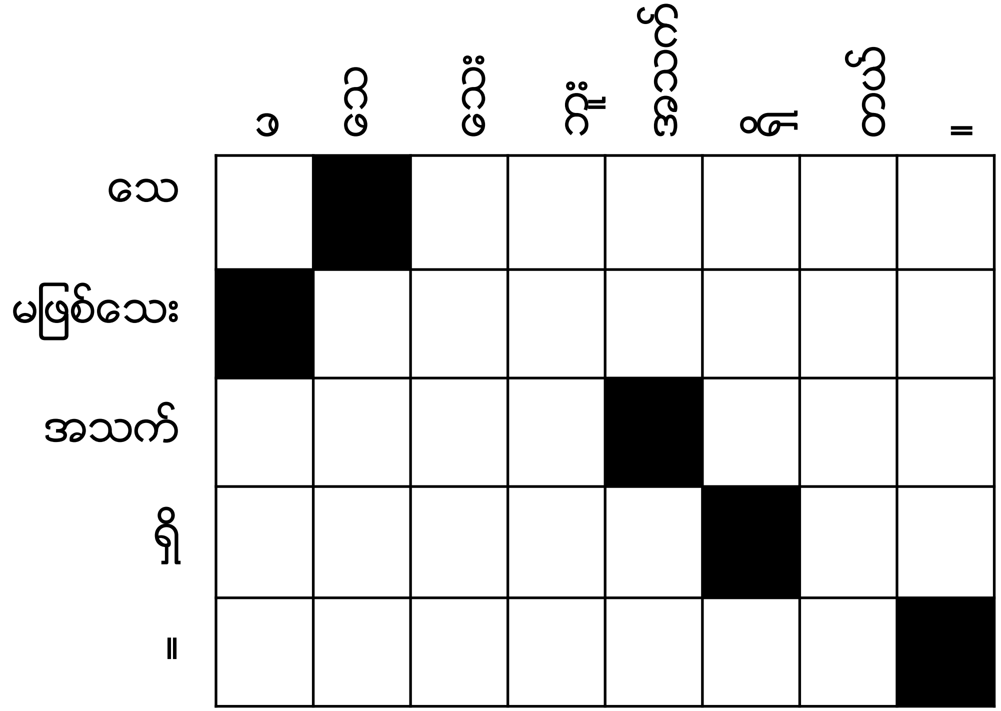
Alignment from P(f | e)



Alignment from P(e | f)

PBSMT (Cont’d)

Finding Alignment Matrices



Intersection of Two Alignments

PBSMT (Cont'd)

Phrase Translation Probability

- For any phrase pair (f,e) extracted from the training data, can calculate

$$\varnothing(f|e) = \frac{count(f,e)}{count(e)}$$

- For example:

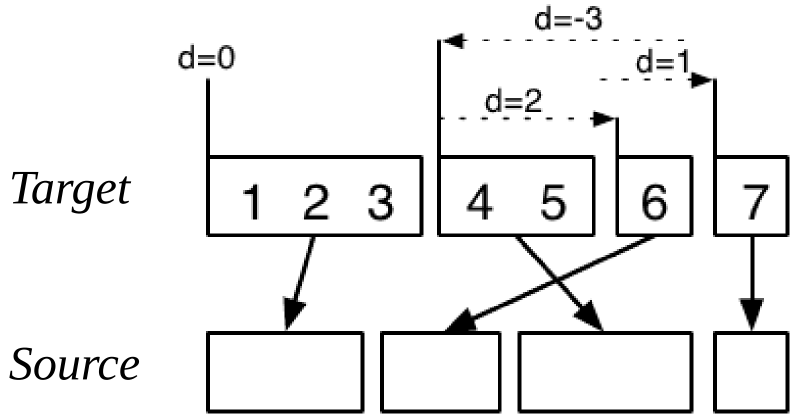
မီးသတ်ဆေးဘူး = 

(အနီ ဘူး ဖြန်း)

$$\varnothing(\text{အနီ ဘူး ဖြန်း} \mid \text{မီးသတ်ဆေးဘူး}) = \frac{count(\text{အနီ ဘူး ဖြန်း , မီးသတ်ဆေးဘူး})}{count(\text{မီးသတ်ဆေးဘူး})}$$

PBSMT (Cont'd)

Distance-based Reordering Model



phrase	translates	movement	distance
1	1–3	start at beginning	0
2	6	skip over 4–5	+2
3	4–5	move back over 4–6	-3
4	7	skip over 6	+1

- distance = $start_i - end_{i-1} - 1$

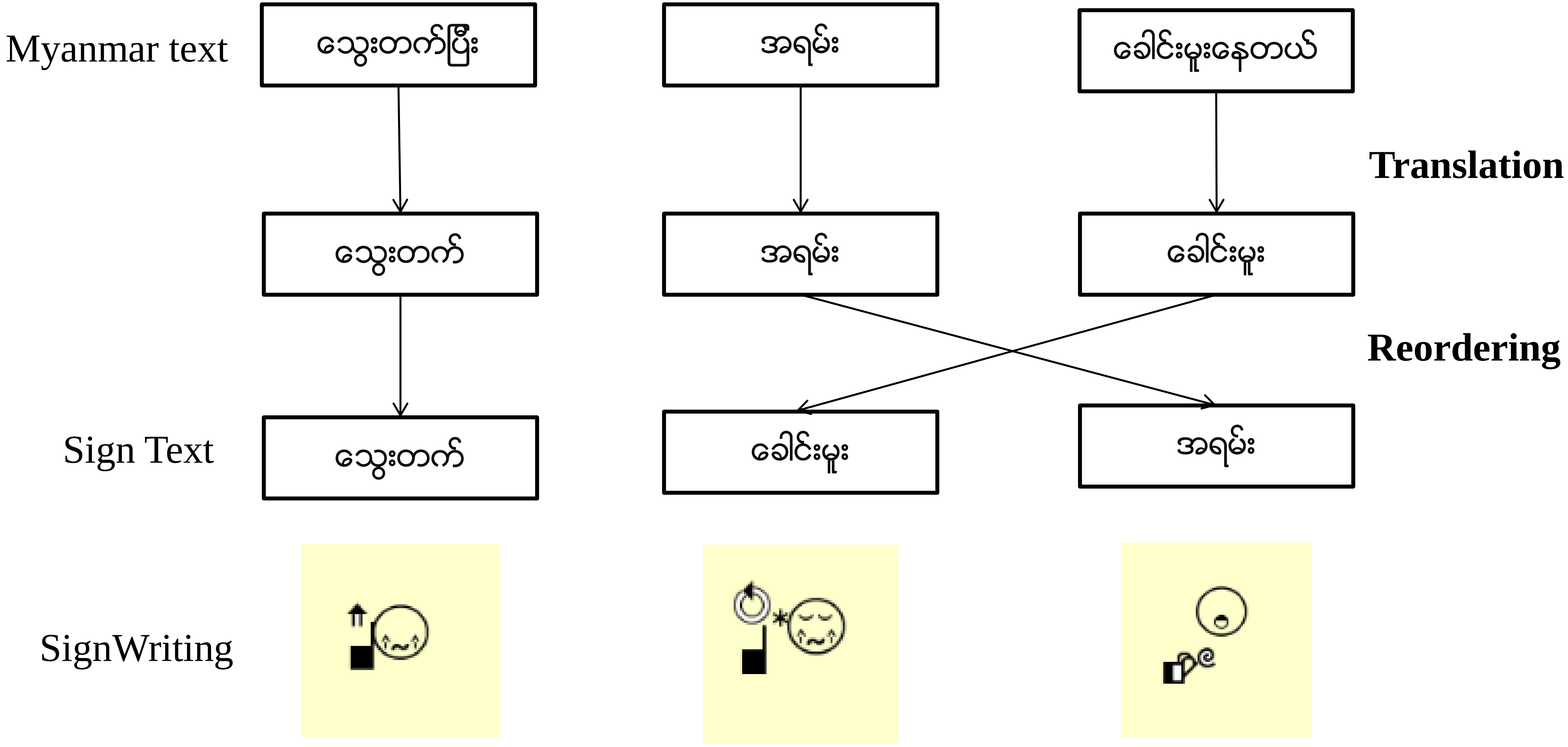
PBSMT (Cont'd)

Example of Phrase Translation Table

- Phrase Translation for "သေ"

Myanmar	*P(e |f)*
သေ	0.102086
မ သေ	0.0113429
သေ ရှိ	0.0510431
သေ ရှိ မရှိ	0.0695603
သေ ရှိ မရှိ ဘာလဲ	0.0113429
သေ ရှိ မရှိ ဘာလဲ ။	0.0510431
….	….

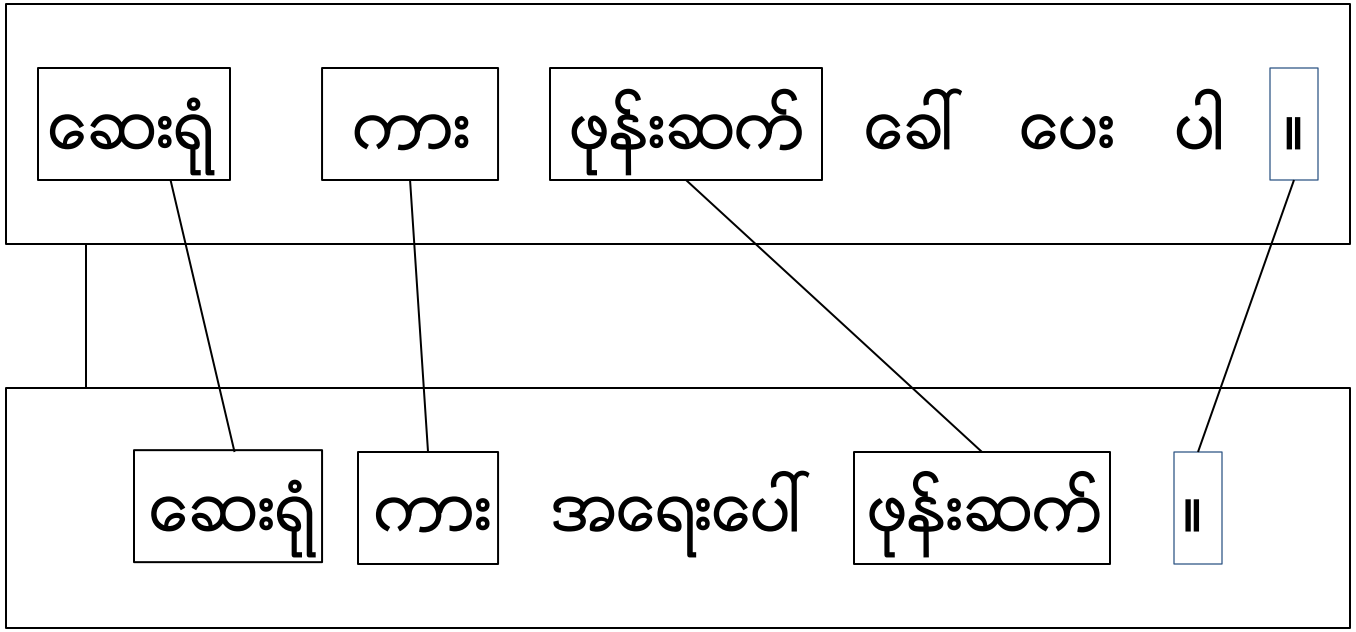
PBSMT (Cont’d)


Myanmar text
သွေးတက်ပြီး
အရမ်း
ခေါင်းမူးနေတယ်
Translation
သွေးတက်
အရမ်း
ခေါင်းမူး
Reordering
Sign Text
သွေးတက်
ခေါင်းမူး
အရမ်း
SignWriting

Hierarchical Phrase-Based Statistical Machine Translation (HPBSMT)

- HPBSMT is a model based on synchronous context-free grammar.
- It learns from a corpus of unannotated parallel text.
- Its advantage over PBSMT is able to represent word reordering process.
- The reordering is represented explicitly rather than encoded into a lexicalized reordering model.
- It is applicable to language pairs that require long-distance re-ordering during translation process.

HPBSMT (Cont’d)



X ⟶ {X1 X2 X3 ခေါ် ပေး ပါ X4 | X1 X2 အရေးပေါ် X3 X4}

Operation Sequence Model (OSM)

- It combines the benefits of phrase-based and N-gram-based SMT and remedies their drawbacks.
- List of Operations can be divided into two groups.
  - **Five Translation Operations**
    - Generate (X,Y),
    - Continue Source Cept,
    - Generate Identical,
    - Generate Source only (X), and
    - Generate Target only (Y).
  - **Three Reordering Operation**
    - Insert Gap,
    - Jump Back (N), and
    - Jump Forward .

Source : ခုချက်ချင်း မီးသတ် ကား လွှတ် လိုက် မယ် ။

Target : မီးသတ် ကား အခု လာမယ် ။

Operations:

o_1 : Insert Gap

Source : ခုချက်ချင်း မီးသတ် ကား လွှတ် လိုက် မယ် ။

Target : မီးသတ် ကား အခု လာမယ် ။

Operations:

o_1 : Insert Gap

o_2 : Generate(မီးသတ်, မီးသတ်)

မီးသတ်

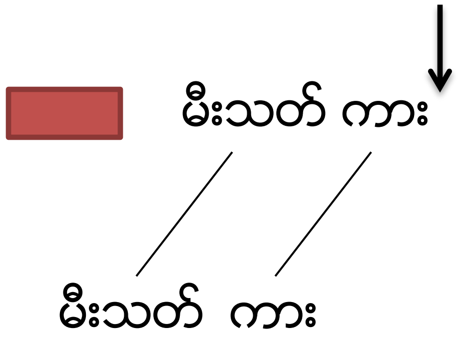
မီးသတ်

Source : ခုချက်ချင်း မီးသတ် ကား လွှတ် လိုက် မယ် ။

Target : မီးသတ် ကား အခု လာမယ် ။

Operations:

o_1 : Insert Gap
o_2: Generate(မီးသတ်, မီးသတ်)
o_3: Generate(ကား, ကား)

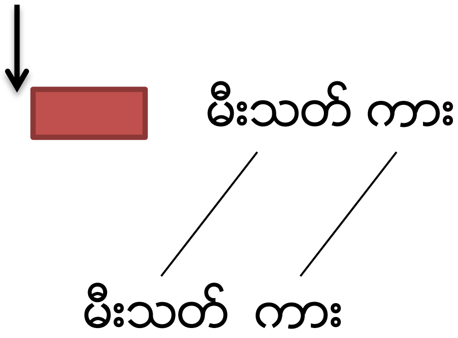


Source : ခုချက်ချင်း မီးသတ် ကား လွှတ် လိုက် မယ် ။

Target : မီးသတ် ကား အခု လာမယ် ။

Operations:

o_1 : Insert Gap
o_2: Generate(မီးသတ်, မီးသတ်)
o_3: Generate(ကား, ကား)
o_4: Jump Back(1)

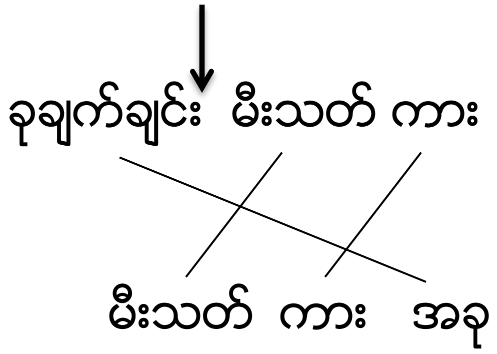


Source : ခုချက်ချင်း မီးသတ် ကား လွှတ် လိုက် မယ် ။

Target : မီးသတ် ကား အခု လာမယ် ။

Operations:

o_1 : Insert Gap
o_2: Generate(မီးသတ်, မီးသတ်)
o_3: Generate(ကား, ကား)
o_4: Jump Back(1)
o_5: Generate(ခုချက်ချင်း, အခု)

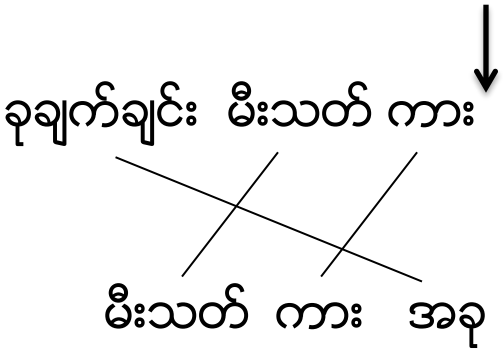


Source : ခုချက်ချင်း မီးသတ် ကား လွှတ် လိုက် မယ် ။

Target : မီးသတ် ကား အခု လာမယ် ။

Operations:

o_1 : Insert Gap
o_2: Generate(မီးသတ်, မီးသတ်)
o_3: Generate(ကား, ကား)
o_4: Jump Back(1)
o_5: Generate(ခုချက်ချင်း, အခု)
o_6: Jump Forward

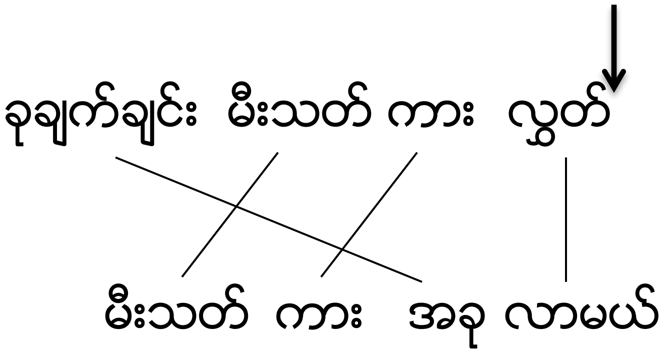


Source : ခုချက်ချင်း မီးသတ် ကား လွှတ် လိုက် မယ် ။

Target : မီးသတ် ကား အခု လာမယ် ။

Operations:

o_1 : Insert Gap
o_2: Generate(မီးသတ်, မီးသတ်)
o_3: Generate(ကား, ကား)
o_4: Jump Back(1)
o_5: Generate(ခုချက်ချင်း, အခု)
o_6: Jump Forward
o_7: Generate(လွှတ်, လာမယ်)

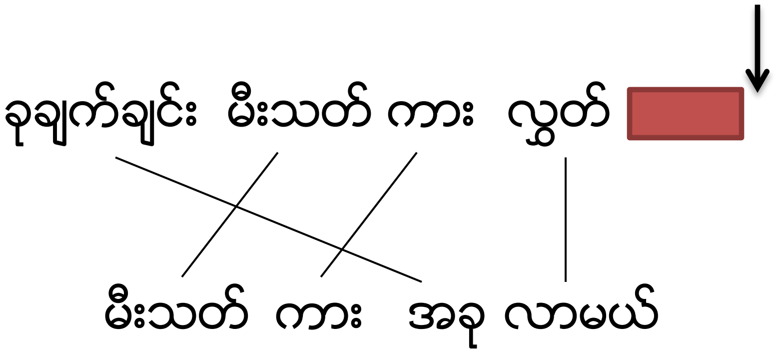


Source : ခုချက်ချင်း မီးသတ် ကား လွှတ် လိုက် မယ် ။

Target : မီးသတ် ကား အခု လာမယ် ။

Operations:

o_1 : Insert Gap
o_2: Generate(မီးသတ်, မီးသတ်)
o_3: Generate(ကား, ကား)
o_4: Jump Back(1)
o_5: Generate(ခုချက်ချင်း, အခု)
o_6: Jump Forward
o_7: Generate(လွှတ်, လာမယ်)
o_8: Insert Gap

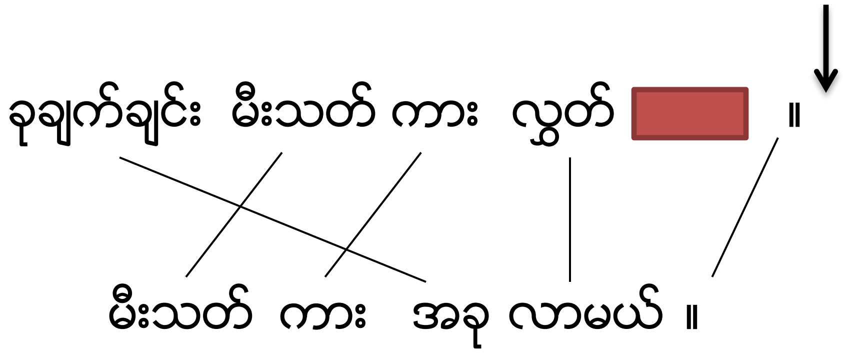


Source : ခုချက်ချင်း မီးသတ် ကား လွှတ် လိုက် မယ် ။

Target : မီးသတ် ကား အခု လာမယ် ။

Operations:

o_1 : Insert Gap
o_2: Generate(မီးသတ်, မီးသတ်)
o_3: Generate(ကား, ကား)
o_4: Jump Back(1)
o_5: Generate(ခုချက်ချင်း, အခု)
o_6: Jump Forward
o_7: Generate(လွှတ်, လာမယ်)
o_8: Insert Gap
o_9: Generate(။, ။)



Evaluation Methods

- Automatic evaluation of machine translation (MT) quality is essential to developing high-quality machine translation systems because human evaluation is time consuming, expensive, and irreproducible.
- If we have a perfect automatic evaluation metric, we can tune our translation system for the metric.
- In NLP, there are many kinds of automatic evaluation methodology (e.g., **BLEU**, NIST, PER, TER, WER, MERT and **RIBES**, etc.).

Evaluation Methods (Cont’d)

- **BLEU –BiLingual Evaluation Understudy**
- It measures how many words overlap in a given translation when compared to a reference translation.
- It computes precision for n-grams of size 1 to 4.
- It ranges from 0-100, the higher the score, the more the translation correlates to a human translation.

Evaluation Methods (Cont'd)

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}.$$

Then,

$$\textsc{Bleu} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n \right).$$

where, c	Hypothesis or Candidate Sentence Length
r	Reference Sentence Length
BP	Brevity Penalty
p	Precision

- Hypothesis: ကျွန်တော့် အစ်ကို ပျောက် သွား တယ် ကို ရှာ ပေး ပါ ။
- Reference: ကျွန်တော့် အစ်ကို ကို ရှာ ပေး ပါ ။
- c = 10, r = 7   and it satisfies c > r so BP = 1
- Use **N=4** and uniform weight $\boldsymbol{w_n = \frac{1}{N}}$

N-gram	$\mathbf{W_n}$	$\mathbf{P_n}$	$\mathbf{Log\ P_n}$	$\mathbf{W_n * log\ P_n}$
1	1/4	7/10	-0.1549	-0.0387
2	1/4	5/9	-0.2553	-0.0638
3	1/4	3/8	-0.426	-0.1065
4	1/4	2/7	-0.5441	-0.136
Total				-0.345
BLEU = 1 * exp (-0.345) = **0.7082 = 70.82%**				

Evaluation Methods (Cont'd)

- **RIBES –Rank-based Intuitive Bilingual Evaluation Score**
- Another failure of BLEU is the **lack** of explicit consideration for **word reordering**.
- To overcome reordering, RIBES uses rank correlation coefficients based on word order to compare SMT and reference translations.
- The primary rank correlation coefficients used are
  - **Spearman's ρ**, which measures **the distance of differences** in rank, and
  - **Kendall's τ**, which measures **the direction of differences** in rank.

Evaluation Methods (Cont'd)

- Spearman's ρ = 1- $\frac{\sum_i {d_i}^2}{{}_{n+1}C_3}$
- Kendall's τ = 2 * $\frac{numbers\ of\ increasing\ pairs}{numbers\ of\ all\ pairs}$ - 1
- Both ρ and τ have the same **range [−1, 1]**.
- These rank measures can be normalized to ensure positive values.
  - Normalized Spearman's ρ (NSR) = (ρ + 1)/2
  - Normalized Kendall's τ (NKT) = (τ + 1)/2
- Kendall's τ is usually smaller values than Spearman's ρ.
- In this experiment, Kendall's τ method is used to calculate RIBES score because of a smaller error sensitivity or more robust and more efficient.

- R0 : ကျွန်တော့် အစ်မ လှေကား ပေါ်က လိမ့်ကျ ဆေးရုံတင် ။
- H0 : ငါ့ လှေကား ပေါ်က ချော်ကျ အစ်မ သွား ။

By removing non-aligned words by one-to-one correspondence,

- R1 : ကျွန်တော့်$_1$ အစ်မ$_2$ လှေကား$_3$ ပေါ်က$_4$ လိမ့်ကျ$_5$ ။$_6$
- H1 : ငါ့$_1$ လှေကား$_3$ ပေါ်က$_4$ ချော်ကျ$_5$ အစ်မ$_2$ ။$_6$
- Word order of R1 : [1, 2, 3, 4, 5, 6]
- Word order of H1 : [1, 3, 4, 5, 2, 6]
- Number of increasing pairs =12
- Number of all pairs = 15
- Kendall's $\tau = 2 * \frac{12}{15} - 1 = 0.6$
- Normalized Kendall's τ (NKT) = (0.6+ 1)/2 = **0.8 = 80%**

Experiment

- This experiment uses a parallel corpus for **Myanmar written text** and **Myanmar SignWriting** in the Emergency domain.
- It contains **888** sentences for emergency situations such as fires, earthquake, floods, storms, accidents, police and health.
  - **550** sentences used for training, **138** sentences used for development and **200** sentences for testing.
- This experiment uses MOSES decoder for machine translation.
- The segmented source and target data are aligned with GIZA++.
- SRILM is used as a language model.

Evaluation Results

Source-Target	BLEU score		
	PBSMT	***HPBSMT***	***OSM***
my-sw	6.28	5.85	**6.31**
sw-my	10.98	11.18	**12.50**

- **BLEU scores for PBSMT, HPBSMT and OSM**

Evaluation Results (Cont’d)

Source-Target	RIBES score		
	PBSMT	***HPBSMT***	***OSM***
my-sw	59.4384	59.6183	**60.5124**
sw-my	65.1887	**66.0532**	62.5728

- **RIBES scores for PBSMT, HPBSMT and OSM**

Conclusion

- In this SMT experiment, testing with Myanmar text and Myanmar SignWriting parallel corpus is finished.
- According to the current SMT experiment results, **OSM** approach gets **highest score** in both **my-sw** and **sw-my** translation measuring with **BLEU** score,
- **OSM** approach gets **highest score** in **my-sw** translation and **HPBSMT** approach gets **highest score** in **sw-my** measuring with RIBES score.
- The future works in this system are to make better performance in three SMT approaches and to make error analysis.

References

- Philipp Koehn., “Statistical Machine Translation: the basic, the novel, and the speculative”
- Nadir Durrani, Helmut Schmid, Alexander Fraser, Philipp Koehn, Hinrich Schütze., “The Operation Sequence Model –Combining N-Gram-based and Phrase-based Statistical Machine Translation”
- Yoav Artzi., “Phrase-based Translation”
- Jörg Tiedemann. Hierarchical phrase-based machine translation
- Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, Hajime Tsukada.,” Automatic Evaluation of Translation Quality for Distant Language Pairs”
- Ameera M.Almasoud and Hend S. Al-Khalifa., “A Proposed Semantic Machine Translation System for translating Arabic text to Arabic sign language”
- Win Pa Pa, Ye Kyaw Thu, Andrew Finch, Eiichiro Sumita,. “A Study of Statistical Machine Translation Methods for Under Resourced Languages”
- Phillip Koehn,. “Statistical Machine Transaltion”
- The population and housing census of Myanmar,2014
- Valerie Sutton; International SignWriting Alphabet (2010)
- Valerie Sutton and Adam Frost; SignWriting Hand Symbols in ISWA2010: Manual 2

Thank you!