UT-YCC

OpU

Statistical Machine Translation from Myanmar Written Text to Myanmar SignWriting (SW)

First 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 November 2017)

Outlines

- Abstract
- Objectives
- Introduction
- Contribution
- Myanmar sign language
- SignWriting (SW)
- Proposed System Design
- Experiment
- 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 represented in form of Myanmar SignWriting.
- There is no Myanmar Written Text and Myanmar SignWriting parallel data yet, and thus we need to prepare it.
- It is based on Statistical Machine Translation (SMT) and Neural Machine Translation (NMT).
- The proposed system will solve difficulties for deaf people to learn the basic concept of daily life, especially in emergency case.

Keywords: *Myanmar SignWriting, Statistical Machine Translation, Neural 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) and Neural Machine Translation (NMT) approaches
- To fulfill the communication requirements between deaf people and hearing people

Introduction

- Sign language - 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
- Limited resources and lack of written language for Myanmar sign language
- I assume Myanmar SignWriting translation system is very important to the Deaf
- My current research will focus on machine translation from Myanmar written text to Myanmar SignWriting

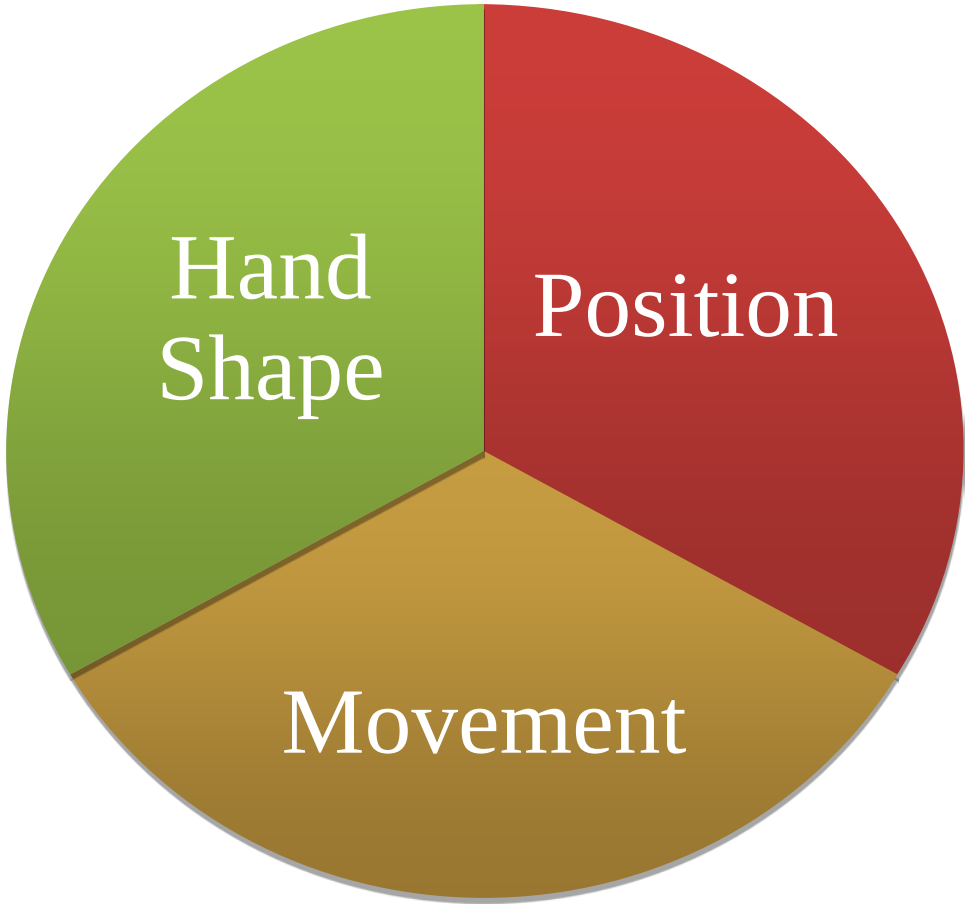
Contribution

- The first study of the Statistical Machine Translation and Neural Machine Translation approaches between Myanmar Written Text and Myanmar SignWriting
- One of my main contribution is to build Myanmar Written Text and Myanmar SignWriting parallel corpus and this will be useful for further researches
- Investigation on statistical and neural machine translation (SMT and NMT) performance on Myanmar written text and Myanmar SignWriting

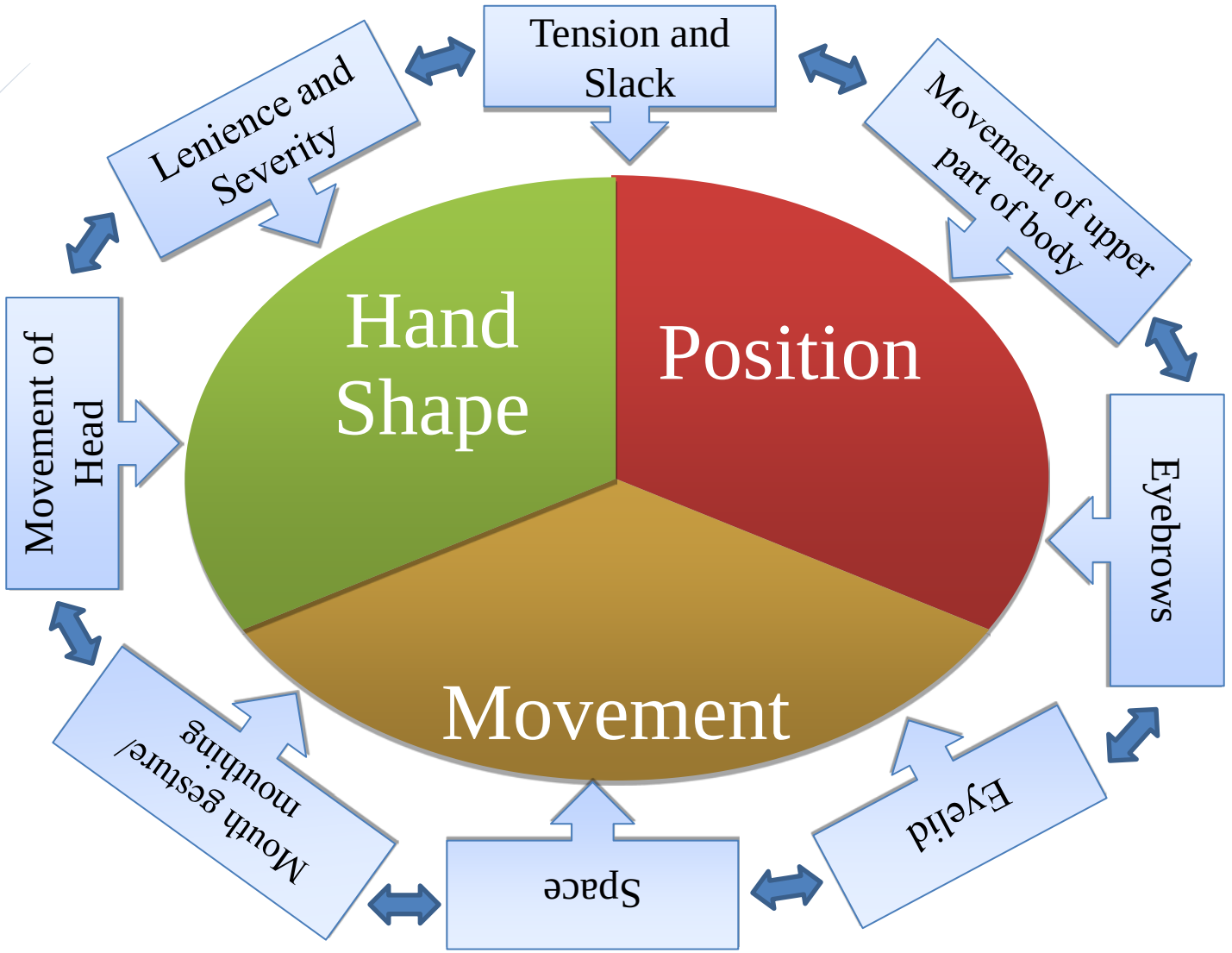
Sign Language

- SL is the native language of the deaf community
- 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).**

Structure of Manual Features (MFs)



Structure of Non-Manual Features (MFs)



Myanmar Sign Language (MSL)

- Each country has its own, native sign language according to their culture
- MSL is a primary communication for Myanmar Deaf community
- There are **four** schools for the Deaf in Myanmar
- Sign language is a vision-based language as they see
- Mandalay and Yangon sign language are difference
- 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
- We need to define Myanmar SignWriting for the Deaf in Myanmar

SignWriting (SW)

- 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: **Handshape** – Index, Index middle, Index middle thumb, four fingers, five fingers, baby fingers, ringer finger, middle finger, index thumb and thumb

Category 2: **Movement** – Contact, Finger Movement, Straight Wall Plane, Straight Diagonal Plane, Straight Floor Plane, Curves Hit Wall Plane, Curves Parallel Wall Plane, Curves Hit Floor Plane, Curves Parallel Floor Plane and Circles

Category 3: Dynamic and Timing – used to give the "feeling" or "tempo" of movement and to show alternating or simultaneous movement.

Category 4: Head and Face – Head, Brow Eyes Eyegaze, Cheeks Ears Nose Breath, Mouth Lips and Tongue Teeth Chin Neck

Category 5: Body – Trunk and Limbs

Category 6: Location – Information of the location of symbols. They are useful for sorting large dictionaries, refining animation

Category 7: Punctuation – used when writing complete sentences or documents in SignWriting

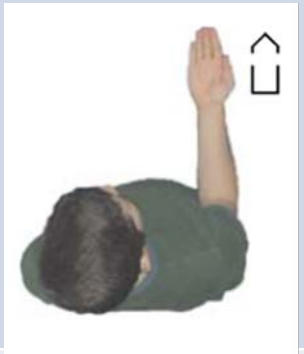
Front View	Top View
	
	
	

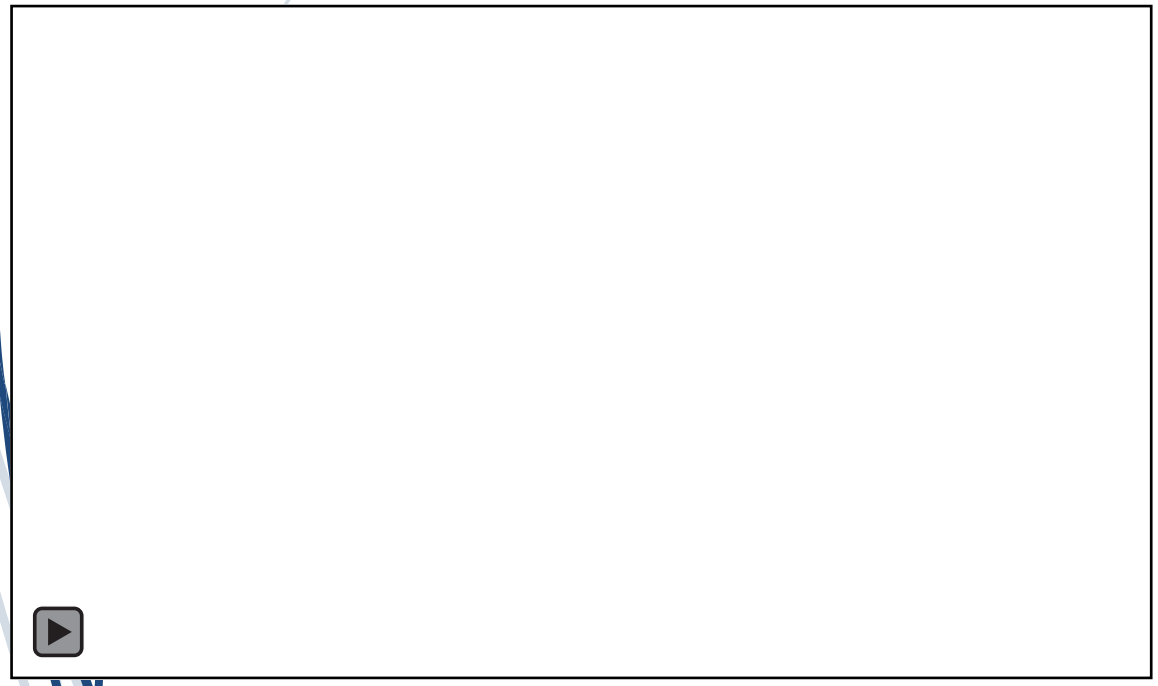
Table 1. Example of SignWriting HAND-FLAT Handshape

Transcription of Myanmar SignWriting

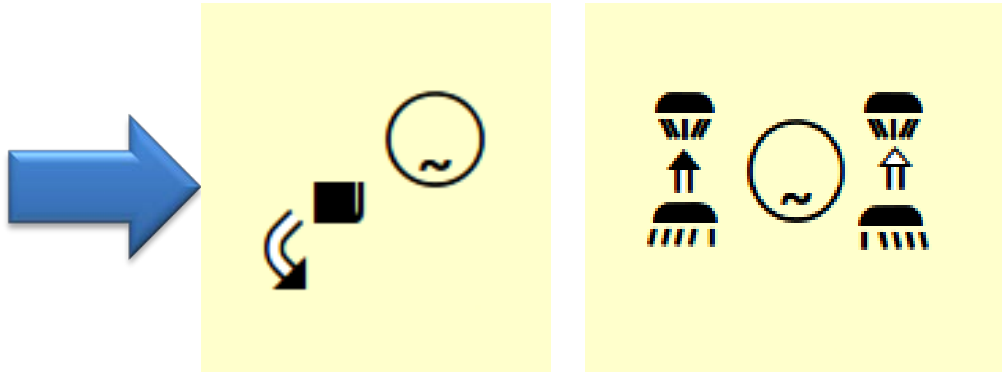
Example 1:

Myanmar Text : မီးခလုတ်ကို ပိတ်ပါ။

Sign Language : မိန်းချ မီးပိတ်။



Sign Language Video



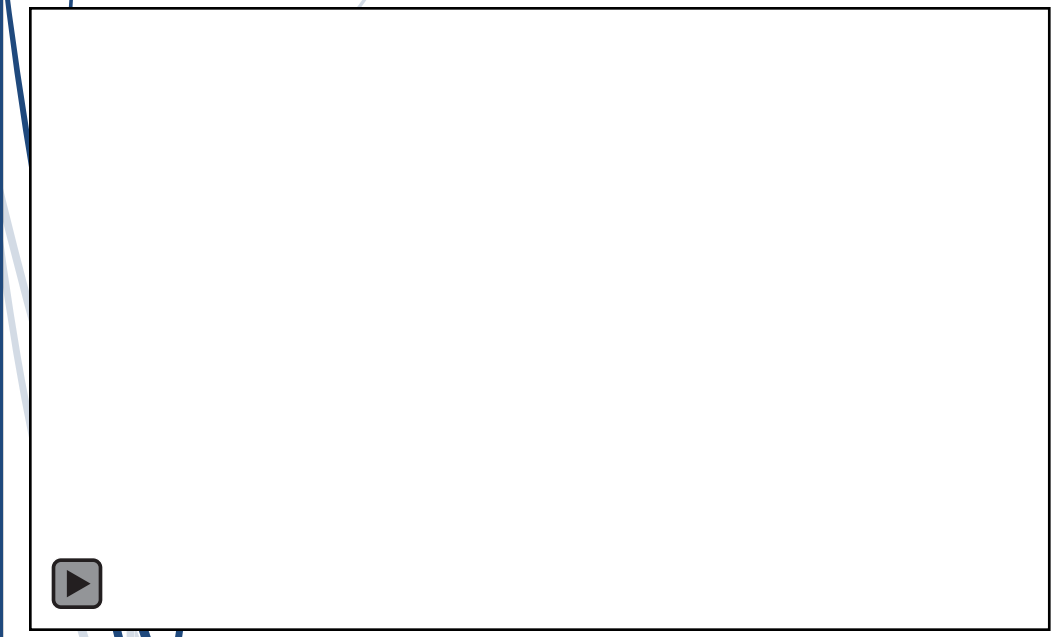
SignWriting

Transcription of Myanmar SignWriting

Example 2:

Myanmar Text : ကျိုချက်ထားသော ရေကို သောက်ပါ။

Sign Language : ရေ ကျိုု သောက် ရ။



Sign Language Video

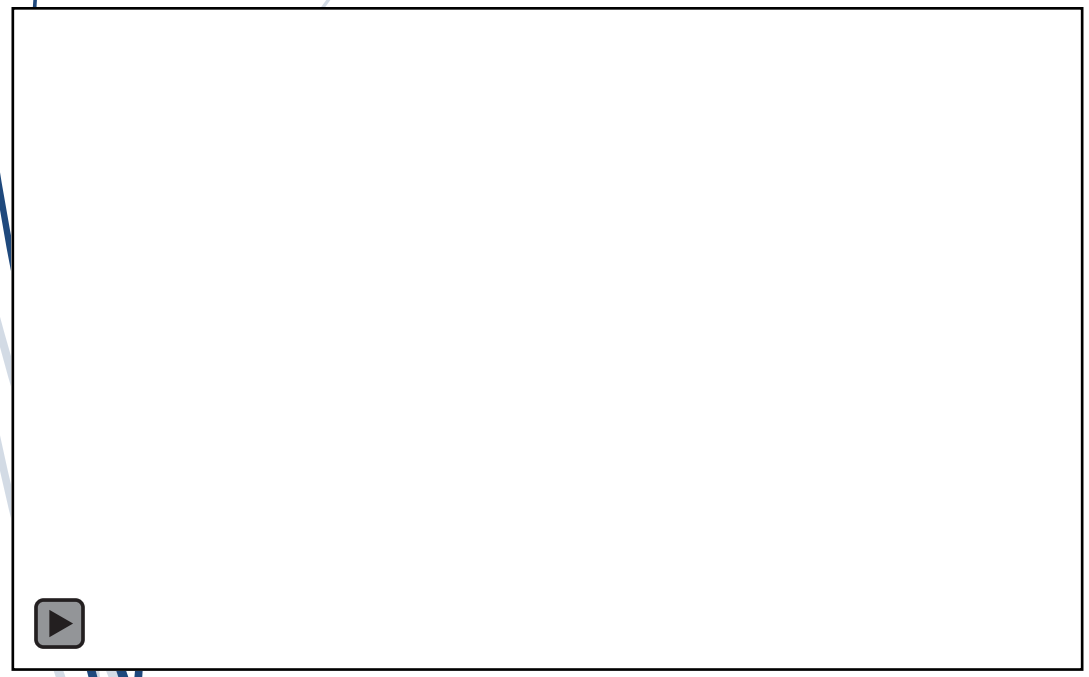
SignWriting

Transcription of Myanmar SignWriting

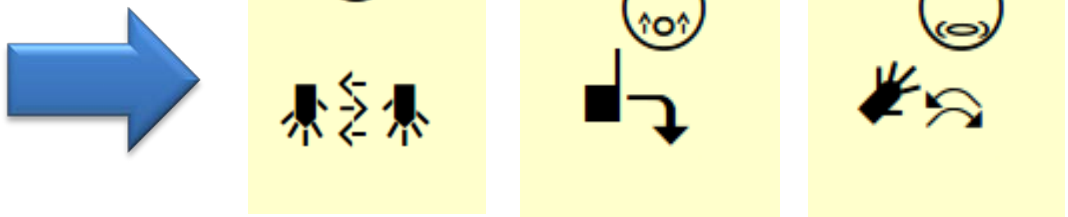
Example 3:

Myanmar Text : ငလျင် ထပ် လှုပ်မယ်။

Sign Language : ငလျင် နောက်ထပ် လာမယ်။



Sign Language Video

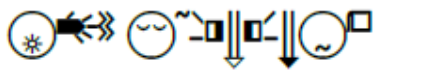
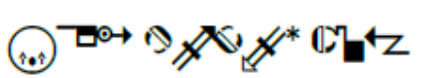
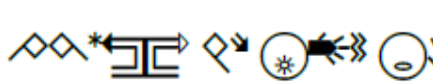
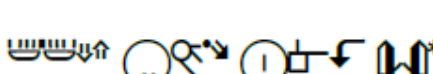
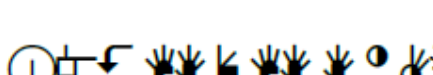
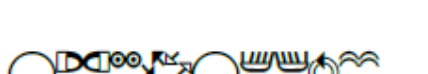


SignWriting

Building Myanmar Written Text and Myanmar SignWriting parallel corpus

- This corpus contains Myanmar written text and the transcribed Myanmar SignWriting.
- There are many **challenges** in building the parallel corpus for SignWriting
  - It does not contain tokenized characters (space, full stop, comma, etc) in Myanmar written text. Tokenization is manual.
  - There is lack of Myanmar sign languages data collected.
  - Learning Myanmar sign language and SignWriting is from zero
  - Myanmar Deaf and Myanmar sign language signers don’t use SignWriting

Building Myanmar Written Text and Myanmar SignWriting parallel corpus

Myanmar Written Text	Myanmar SignWriting
မီး ငြိမ်း ။	
မီးချိတ် ။	
မီးကပ် ။	
ရေ ပုံး ။	
သဲ အိတ် ။	
မီးသတ်ဆေးဘူး ။	
မီးသတ်ရေကန် ။	
မီးလောင် လွယ် သော ပစ္စည်း များ ။	
လောင်စာဆီ ။	
အမှိုက် ။	

Statistical Machine Translation (SMT)

- 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
- The proposed system will use three SMT approaches:
  - Phrase-Based Statistical Machine Translation (PBSMT),
  - Hierarchical Phrase-Based Statistical Machine Translation (HPBSMT) and
  - Operation Sequence Model (OSM).

Statistical Machine Translation (SMT)

- Let f be any text in the source language.
- The translation ê is searched among texts in the target language e using the probabilistic model:

$$\hat{e} = \arg max_e P(f \mid e) \mid P(e)$$

  - Translation Model : P(f |e)
  - Language Model : P (e)
- Decomposition of Translation Model

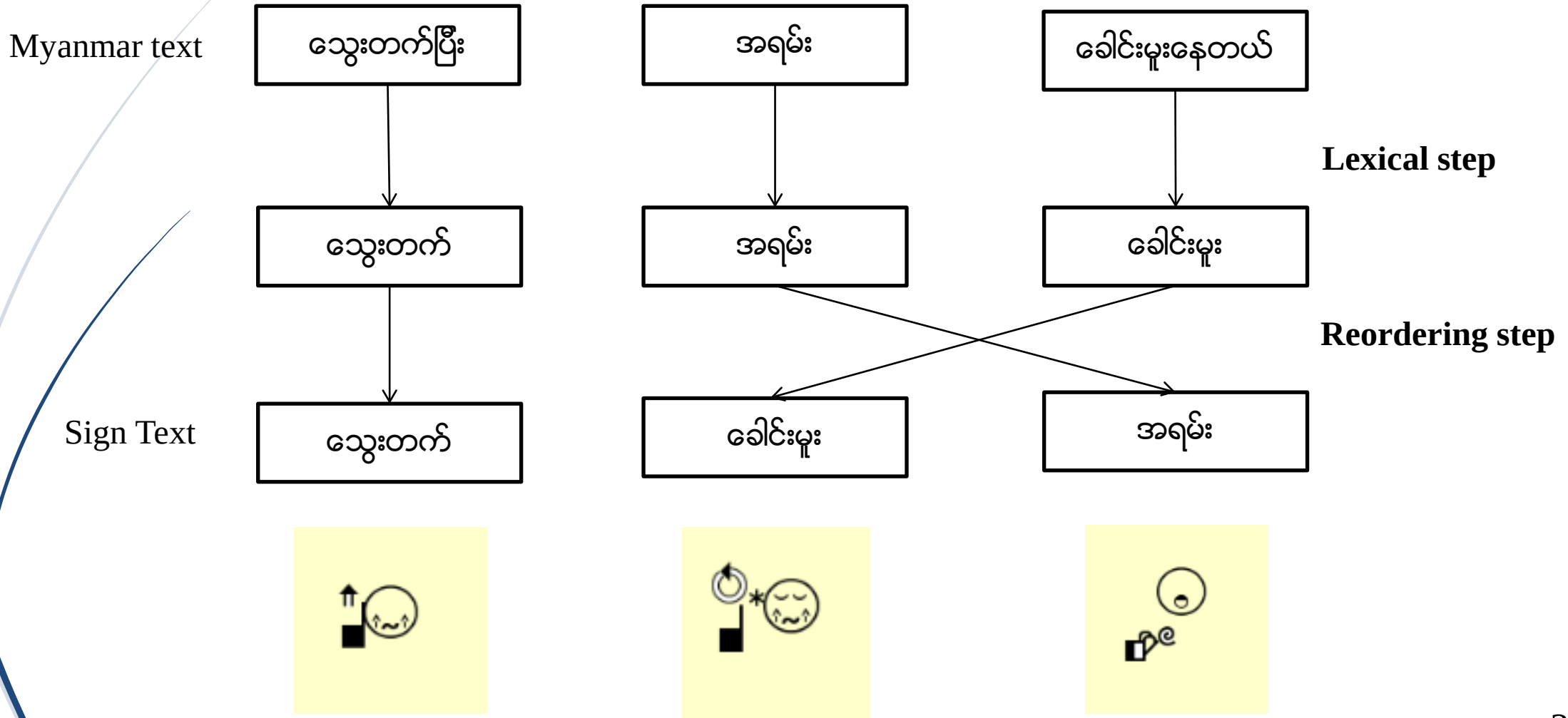
$$e_{best} = \text{argmax}_e \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i)\, d(start_i - end_{i-1} - 1)\, p_{LM}(e)$$

  - Phrase translation probability : Ø
  - Reordering probability : d

Phrase-Based Statistical Machine Translation (PBSMT)

- PBSMT translate phrases as atomic units
  - A continuous sequence of words
  - Not necessarily a linguistic phrase
- Better translation performance than word-based
- Consist of
  - Phrase-pair probabilities extracted from corpus
  - Reordering model
  - An algorithm to extract the phrases to build a phrase-table

Phrase-Based Statistical Machine Translation (PBSMT)



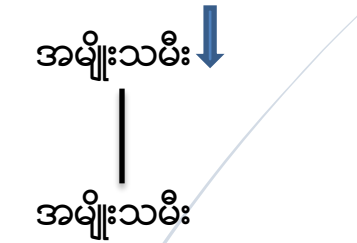
Hierarchical Phrase-Based Statistical Machine Translation (HPBSMT)

- Model based on synchronous context-free grammar
- Learn from a corpus of unannotated parallel text
- Its advantage over PBSMT is able to represent and word reordering process
- It is applicable to language pairs that require long-distance re-ordering during translation process

Operation Sequence Model (OSM)

- 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)
    - Generate Target only (Y)
  - Three Reordering Operation
    - Insert Gap
    - Jump Back (N)
    - Jump Forward

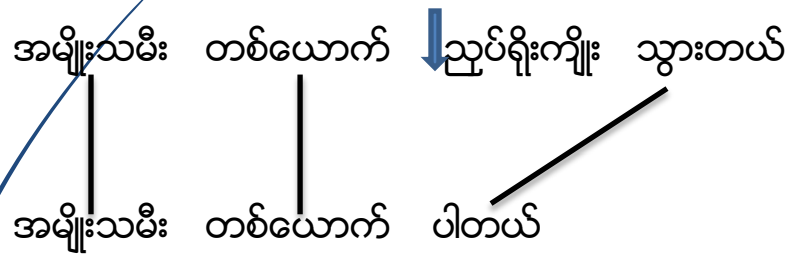
Example of Operation Sequence Translation



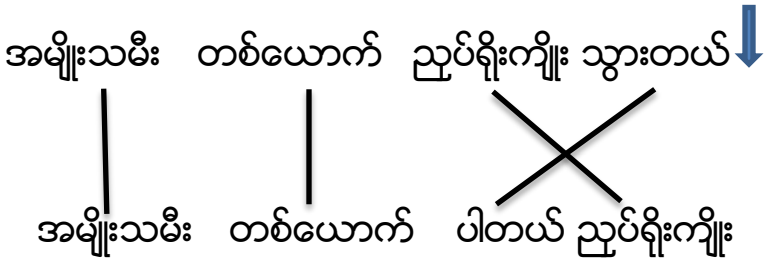
1.Generate
(အမျိုးသမီး,အမျိုးသမီး)

2.Generate
(တစ်ယောက်,တစ်ယောက်)

3.Insert Gap, Generate (သွားတယ်, ပါတယ်)



4.Jump Back (2)



5.Generate (ညှပ်ရိုးကျိုး, ညှပ်ရိုးကျိုး)

Neural Machine Translation (NMT)

- An approach to machine translation that uses a large neural network
- It departs from phrase-based statistical translation approaches.
- Google and Microsoft translation services now use NMT.
- NMT models use deep learning and representation learning.
- All parts of the neural translation model are trained jointly (end-to-end) to maximize the translation performance.

System Design

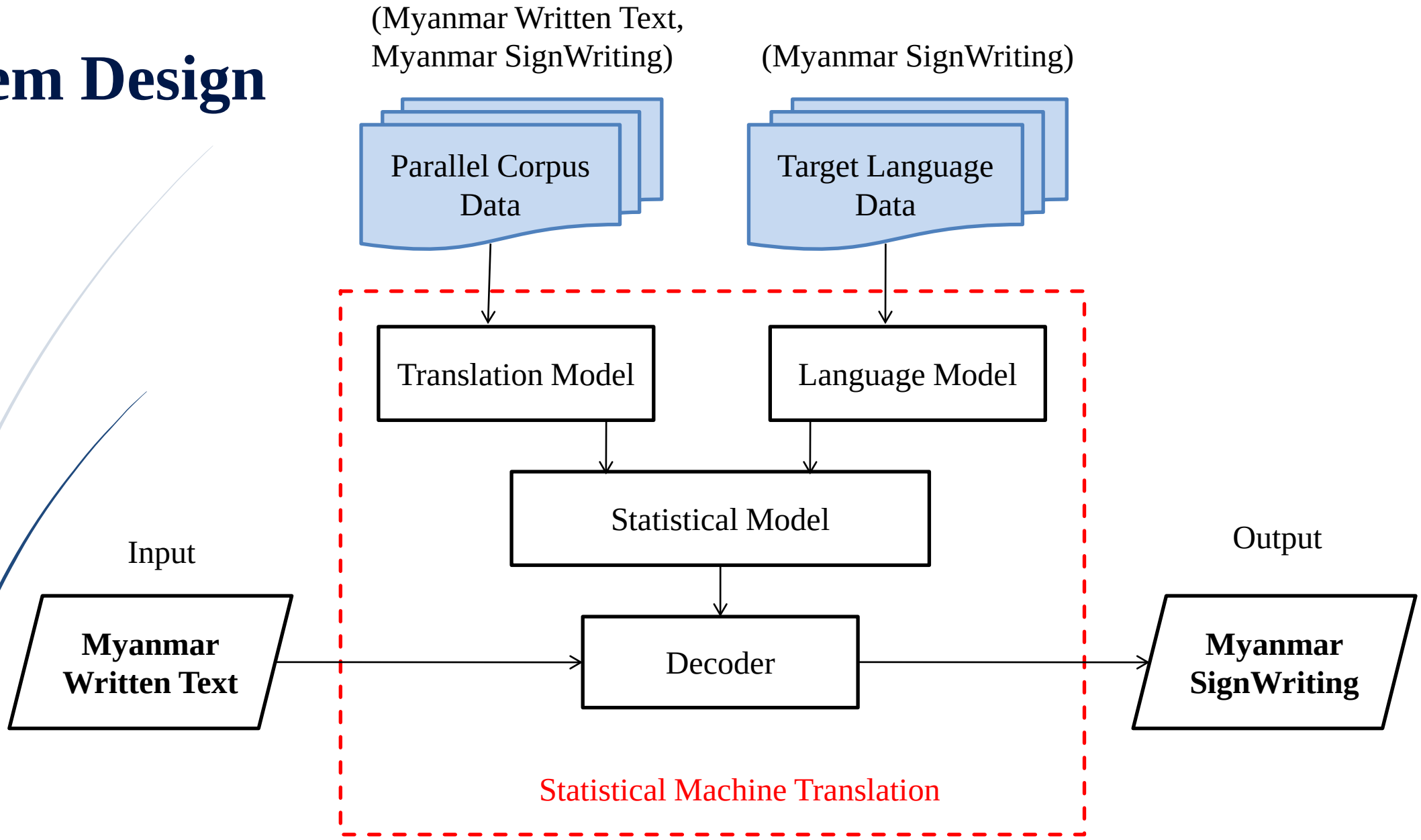


Figure : Flow diagram of the proposed system

Experiment

- Use a parallel corpus for **Myanmar written text** and **Myanmar sign language text** in the Emergency domain
- Contain 558 sentences for emergency situations such as fires, earthquake, floods, storms, accidents, police and health
- 400 sentences used for training, 100 sentences used for development and 58 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.

Conclusion

- In example SMT translation, testing with Myanmar text and Sign Text parallel corpus is finished
- The future works in this proposed system are:
  - Preparing Myanmar SignWriting data
  - Translating with Myanmar Written Text and Myanmar SignWriting
  - Evaluating results

References

- Cayley Guimaraes, Jeferson F. Guardezi, Luis Eduardo Oliveira, Sueli Fernandes ., “Deaf Cultures and Sign Language Writing System – a Database for a New Approach to Writing System Recognition Technology”
- Sara Morrissey and Andy Way., “Joining Hands: Developing a Sign Language Machine Translation System With and For the Deaf Community”
- 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”
- https://en.wikipedia.org/wiki/Statistical_machine_translation
- https://en.wikipedia.org/wiki/Neural_machine_translation
- 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.