UT-YCC

OpU

Statistical Machine Translation between Myanmar Written Text, Myanmar Sign Language and Myanmar SignWriting

Third Seminar

Presented by: Hnin Wai Wai Hlaing (ME-IST 23)

Supervisor: Dr. Nandar Win Min

Co-supervisor: Dr. Ye Kyaw Thu, Daw Swe Zin Moe

(17th May 2018)

Outlines

- Abstract
- Objectives
- Motivation
- Sign Language
- SignWriting
- Building parallel corpus
- Segmentation
- Experiments and Analysis
- Conclusion

Abstract

- In the field of Machine Translation, significant progress has been made by using statistical methods.
- The proposed system suggests a **statistical machine translation between Myanmar Written Text, Myanmar Sign Language and Myanmar SignWriting**.
- Developing MWT, MSL, MSW parallel corpus is used for the translation.
- The corpus is only focused on the **Emergency domain** (e.g., fires, earthquake, floods, storms, accidents, police, health, number, date and time, etc).
- Three translation experiments are done: MWT-MSW, MWT-MSL, MSL-MSW and vice versa.

Abstract (Cont'd)

- Experiments were carried out using three different statistical machine translation (SMT) approaches:
  - ✓ **phrase-based statistical machine translation (PBSMT)**
  - ✓ **hierarchical phrase-based statistical machine translation (HPBSMT)**
  - ✓ **operation sequence model (OSM)**
- Three different segmentation were applied:
  - ✓ **Syllable segmentation for MWT and MSL,**
  - ✓ **Word segmentation for MWT and**
  - ✓ **Sign unit based segmentation for MSL.**
- The system can solve difficulties for Deaf in daily life and helps them especially **emergency situations**.

Objectives

- To learn Machine Translation between Myanmar Written Text, Myanmar Sign Language and Myanmar SignWriting
- To develop MWT, MSL, MSW parallel data corpus
- To measure Machine Translation performance using Statistical Machine Translation (SMT) approaches
- To introduce SignWriting to the Myanmar Deaf society
- To fulfill the communication requirements between Deaf and hearing people

Motivation

- According to 2014 Myanmar National census, about 1.3 percent of population in Myanmar are Deaf or hard-of-hearing people.
- They face various difficulties in communicating with other hearing people and feel isolating from their surroundings.
- They are limited resources of information written in their language.
- The main motivation is to introduce SignWriting to the Myanmar Deaf society using statistical machine translation.
- This will reduce the gap between the Deaf and hearing people of the country.

Sign Language (SL)

- SL is the native language of the Deaf community.
- It is a Vision-based language as Deaf can see.
- Deaf 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.
- People who used sign language are:
  - ✓ Deaf people,
  - ✓ Hard-of-hearing people and
  - ✓ People who cannot speak.

Manual Features(MFs)

- Hand Shape
- Position
- Movement

Non-Manual Features(NMFs)

- Movement of Head
- Tension and Slack
- Movement of upper part of body
- Eyebrows
- Eyelid
- Mouth gesture

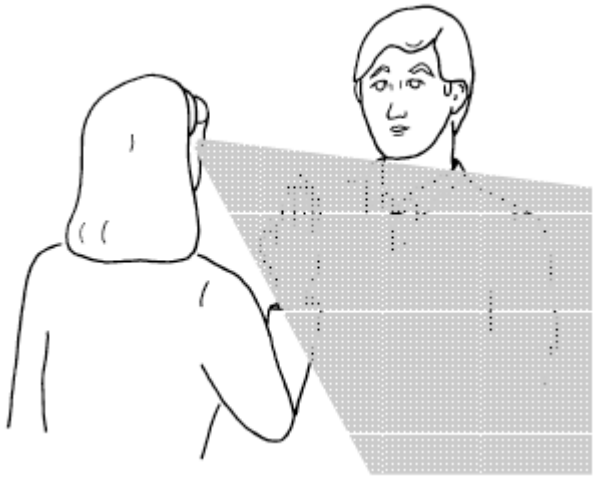
Figure 1 : Structure of Sign Language

Myanmar Sign Language (MSL)

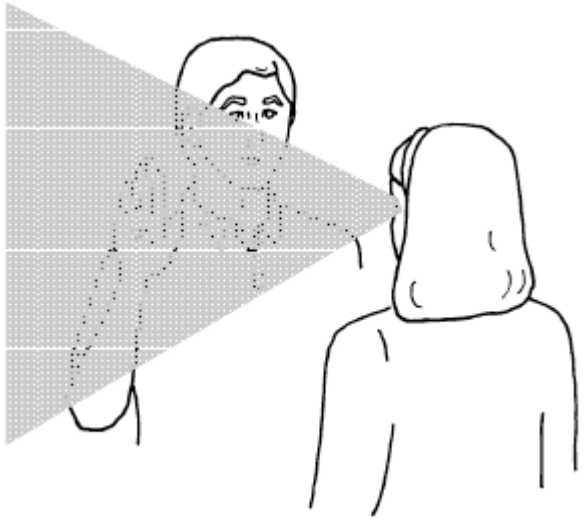
- SL is not a universal language –each country has its own, native sign language according to their culture.
- There are four schools for the Deaf in Myanmar.
- MSL is a primary communication for Myanmar Deaf community.
- It has its own grammar structure which is very difference with Myanmar written text.
- A number of written systems for representing sign languages have been developed, and SignWriting Alphabets are defined for each sign in each country.
- Myanmar SignWriting for the Deaf is needed to define for each sign.

SignWriting (SW)

- SW was developed in 1974 by Valerie Sutton, dancer and movement analyst.
- It is a writing system that is a sequence of symbols for sign language.
- Deaf represents two perspective: **signer's perspective** and **observer's perspective.**



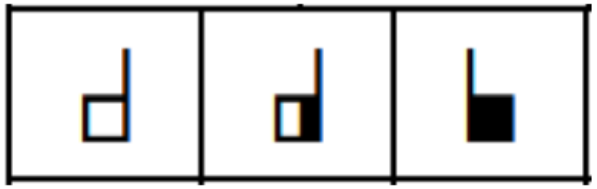
Signer's Perspective



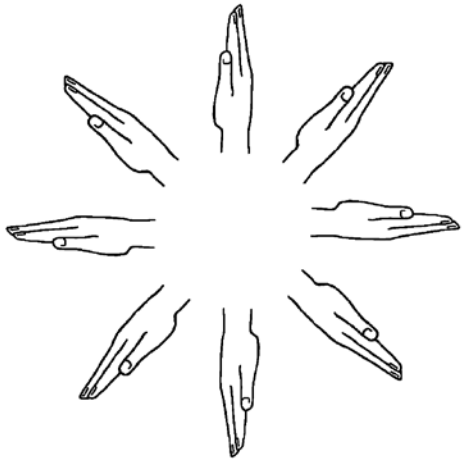
Observer's Perspective

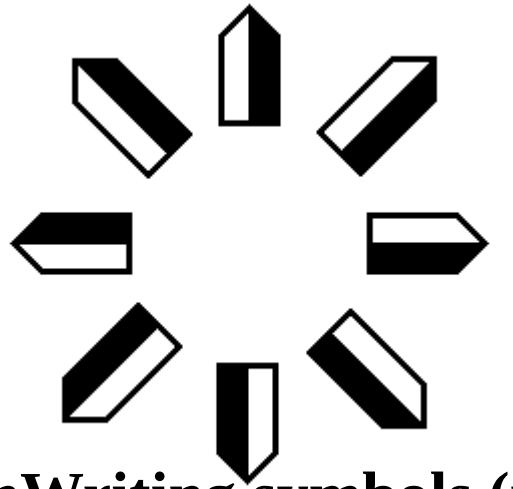
SignWriting (Cont'd)

- It is written horizontally (**left to right**) and the **right hand is dominant**.
- Hand orientation is important for SignWriting and there are 3 different filling symbols and 8 different spatial rotations symbols for each hand.
- **International Sign Writing Alphabet (ISWA) 2010** defines 7 categories and 30 groups to form the base symbols.



(a)Three filling symbols





(b) 8 rotation of hand (left) and SignWriting symbols (right)

SignWriting (Cont'd)

Front View

Top View

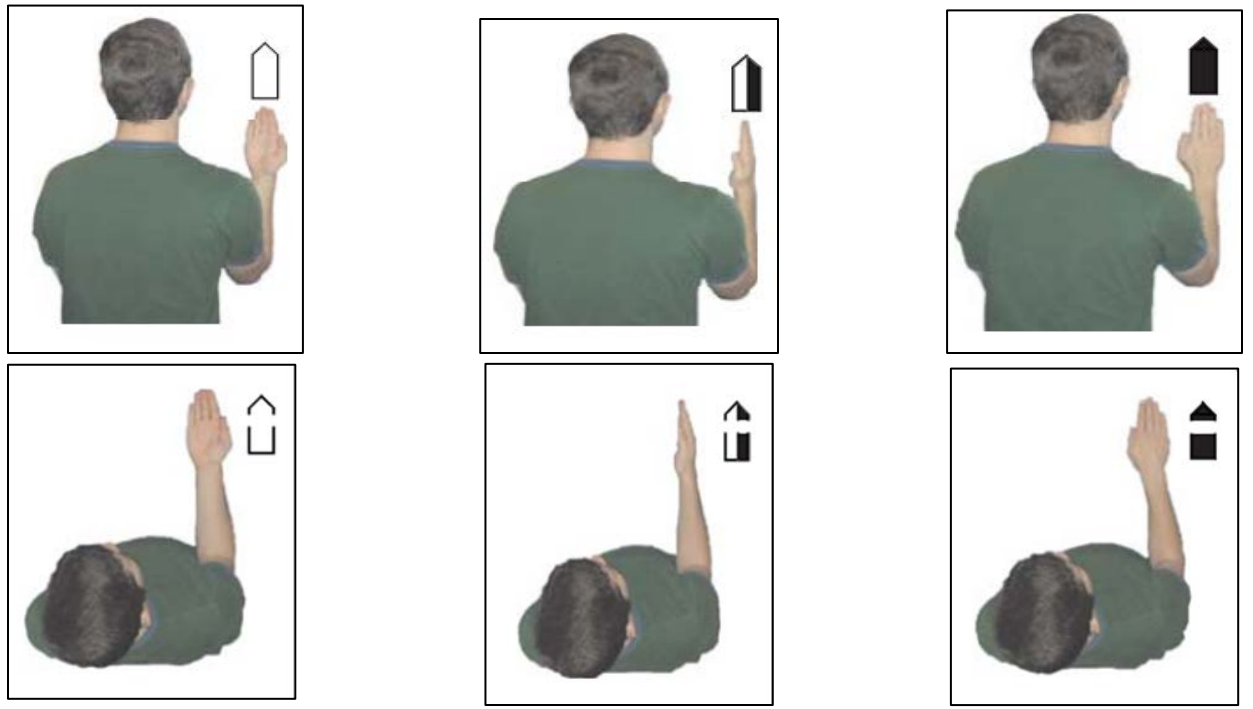
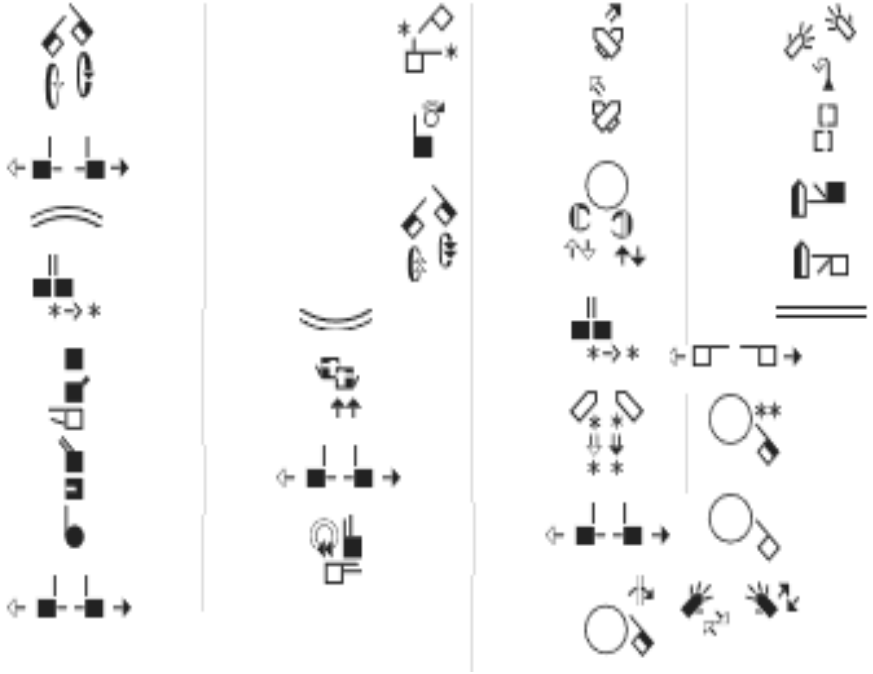


Figure 2: Example of SignWriting HAND-FLAT hand shapes

Building Parallel Corpus

- The corpus contains Myanmar Written Text, transcribed Myanmar Sign Language and the Myanmar SignWriting.
- There are many challenges in building the parallel corpus for SignWriting.
  - ✓ There is lack of Myanmar sign languages video data collected.
  - ✓ It is difficult to understand sign language and its usage.
  - ✓ SignWriting symbols need to clearly define for each Sign.
  - ✓ It does not widely use SignWriting in Myanmar Deaf Community.

Building Parallel Corpus(Cont’d)



- **Video Data Collection**
- **Manual Annotation with SignWriting**

Building Parallel Corpus(Cont'd)

Video Data Collection

- **Selection of Myanmar Sentence**

Building Parallel Corpus(Cont'd)

Video Data Collection

- Selection of Myanmar Sentence
- **Discussion with Sign Language Trainers and Native Signers**

Building Parallel Corpus(Cont'd)

Video Data Collection

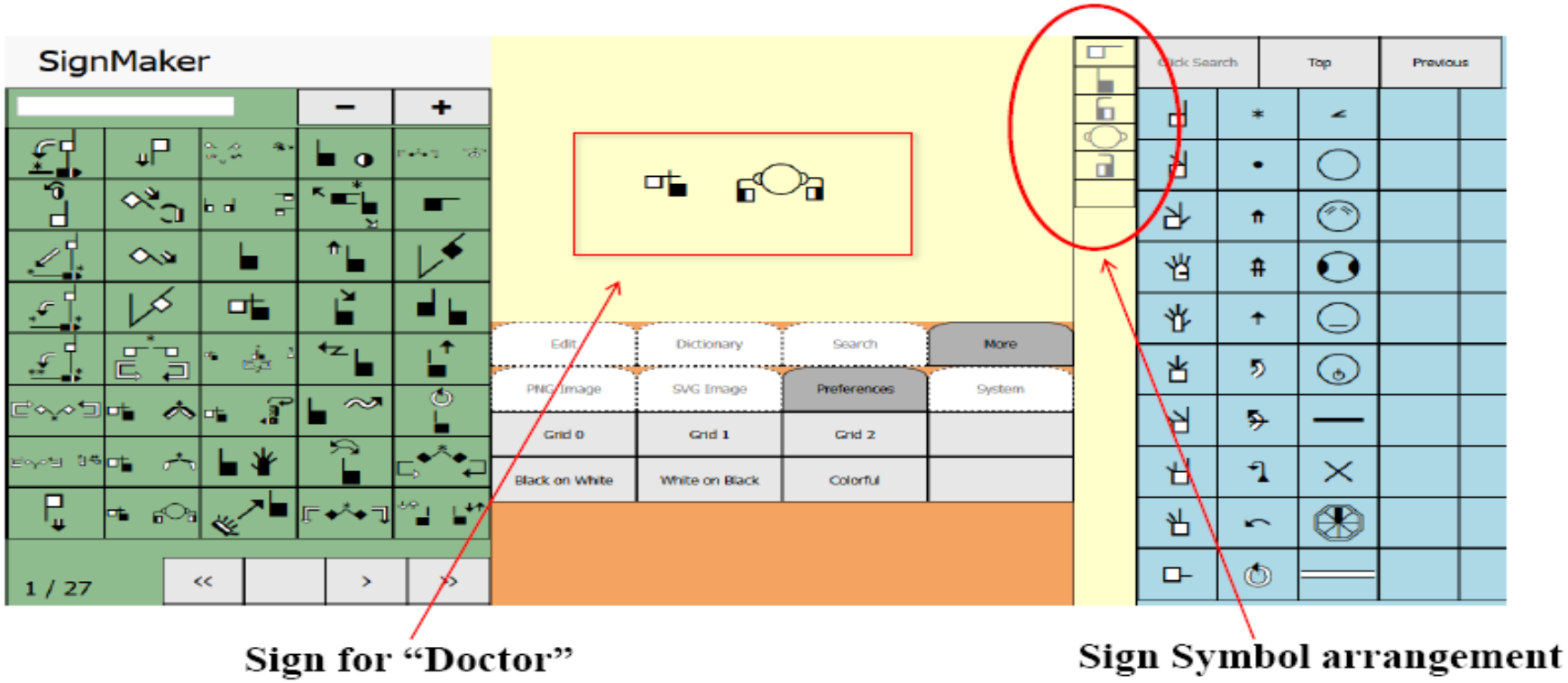
- Selection of Myanmar Sentence
- Discussion with Sign Language Trainers and Native Signers
- **Video Recording**

Building Parallel Corpus(Cont'd)

Manual Annotation with SignWriting

- Looking each sign in video (Non-manual and Manual Signs)

Building Parallel Corpus(Cont'd)



Manual Annotation with SignWriting

- Looking each sign in video (Non-manual and Manual Signs)
- **Sign symbol arrangement in SignMaker**

Building Parallel Corpus(Cont'd)

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

Manual Annotation with SignWriting

- Looking each sign in video (Non-manual and Manual Signs)
- Sign symbol arrangement in SignMaker
- **Defining Unicode sequence and Converting SignWriting symbols**

MWT-MSW Parallel Corpus Data

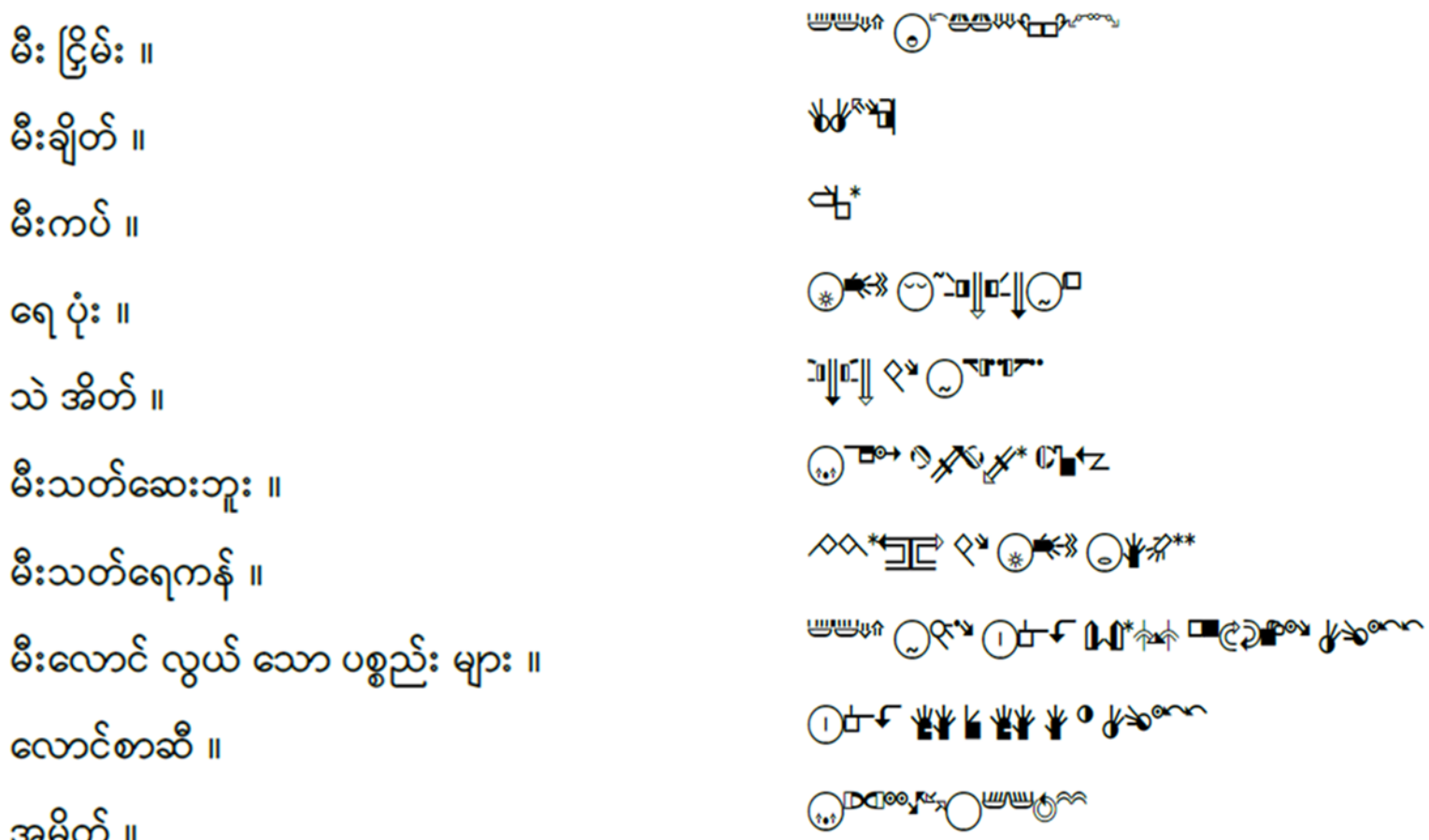


Figure 3 : Example of MWT-MSW parallel corpus

Segmentation

- In SMT, word segmentation is a necessary step in order to yield a set of tokens upon which the alignment and indeed the whole learning process can operate.
- Based on the previous studies of Myanmar word segmentation schemes, three segmentation schemes are used for MWT and MSL.

Syllable Segmentation for MWT and MSL

- Basic units for pronunciation of Myanmar words
- Consonant based syllables –describes with Backus Normal Form(BNF) as follows:
  - ✓ Syllable := C{M}{V}[CK][D]

  C for consonants, M for medials, V for vowels, K for vowel killer character and D for diacritic character
- Myanmar syllable segmentation can be done with rule-based, finite state automation(FSA) or regular expression (RE).
- This work use RE based Myanmar syllable segmentation tool, named “sylbreak”

https://github.com/ye-kyaw-thu/sylbreak

Sign Unit Based Segmentation for MSL

- For MSL sentences, segmentation is based on meaningful MSL word other sign languages such as ASL, BSL and JSL.
- Some examples of MSL word category are:
  - ✓ Repeated sign(e.g. two or more repeated “thank you” sign for “please”)
  - ✓ Sign with multiple meanings(e.g. one MSL sign for “blood” and “red”)
  - ✓ Compound sign (e.g. combination of MSL signs “car”, “emergency” and “fire extinguishing” for “fire truck”
  - ✓ Name sign (e.g. Pyin Oo Lwin City)
  - ✓ Fingerspelling sign(e.g. “O” sign + “2” sign for “O_2”)
  - ✓ Fingerspelling shortcut sign (“O” for Octane, Myanmar consonant “မ” (Ma) for Mandalay city)
  - ✓ Phrase or sentence level signs (e.g. MSL sign for စိတ်ငြိမ်ငြိမ်ထား(calm down) , ကားတိုက်(car accident))

Experiment

(1) Corpus Statistics

- The experiment uses a parallel corpus for **Myanmar written text, Myanmar SignWriting** and **Myanmar SignWriting** in **Emergency domain**.
- Current parallel corpus size is limited and 10-fold cross validation is done for all experiments.
- Total sentences in the corpus contain 888 sentences**.**
  - ✓ **600 sentences** are used for training.
  - ✓ **138 sentences** are used for development.
  - ✓ **150 sentences** are used for testing.

Figure 3: Cross Validation Strategy

Experiment

(2)Moses SMT System

- It uses PBSMT, HBPSMT, OSM system provided by the Moses toolkit.
- Word segmented source language are aligned with the word segmented target language using GIZA++.
- KENLM is used as a language model.

Evaluation Results

src-trg	MWT (word)-MSW		
	PBSMT	***HPBSMT***	***OSM***
my-sw	**12.476** (0.6526)	12.328 (0.6525)	12.427 **(0.6498)**
sw-my	22.268 (0.7184)	22.209 (0.7281)	**23.063 (0.7295)**

src-trg	MWT(syllable)-MSW		
	PBSMT	***HPBSMT***	***OSM***
my-sw	9.108 (0.6358)	**9.432 (0.6388)**	9.273 (0.6289)
sw-my	20.345 (0.7231)	20.081 (0.7234)	**21.628 (0.7288)**

*the values inside the parentheses are RIBES scores

Table 1: BLEU and RIBES scores of MWT and MSW pairs for PBSMT, HPBSMT and OSM using 888 sentences

Evaluation Results

src-trg	MWT (word)-MSL(word)		
	PBSMT	***HPBSMT***	***OSM***
my-sl	16.62 (0.7698)	**17.50 (0.7781)**	16.24 (0.7752)
sl-my	**21.64 (0.7881)**	20.80 (0.7881)	20.84 (0.7859)

src-trg	MSL(syllable)-MWT(word)		
	PBSMT	***HPBSMT***	***OSM***
my-sl	33.98 (0.8132)	35.17 (0.8165)	**35.25 (0.8225)**
sl-my	32.63 (0.7834)	20.80 (0.7881)	**20.84 (0.7859)**

*the values inside the parentheses are RIBES scores

Table 2: BLEU and RIBES scores of MWT and MSL pairs for PBSMT, HPBSMT and OSM using 888 sentences

Evaluation Results

src-trg	MSL(word)-MSW		
	PBSMT	***HPBSMT***	***OSM***
sl-sw	49.25 (0.8527)	47.78 (0.8541)	**49.93 (0.8428)**
sw-sl	46.83 (0.8543)	46.72 (0.8541)	**47.21 (0.8532)**

src-trg	MSW-MSL(syllable)		
	PBSMT	***HPBSMT***	***OSM***
sl-sw	33.14 (0.8218)	32.98 (0.8038)	**34.09 (0.8330)**
sw-sl	**50.78** (0.8817)	49.69 (0.8804)	50.27 **(0.8836)**

*the values inside the parentheses are RIBES scores

Table 3: BLEU and RIBES scores of MSL and MSW pairs for PBSMT, HPBSMT and OSM using 888 sentences

Experiment

(3)Additional experiments

- The experiments is tested with bigger size corpus.
- Total sentences in the corpus contain 1448 sentences.
  - ✓ 1000 sentences are used for training.
  - ✓ 170 sentences are used for development.
  - ✓ 278 sentences are used for testing.

Evaluation Results

src-trg	MWT (word)-MSW		
	PBSMT	***HPBSMT***	***OSM***
my-sw	8.39 (0.6183)	**8.65** (0.6186)	8.33 **(0.6095)**
sw-my	**11.23** (0.6453)	10.95 **(0.6549)**	11.11 (0.6539)

src-trg	MWT(syllable)-MSW		
	PBSMT	***HPBSMT***	***OSM***
my-sw	7.46 (0.6161)	**7.99 (0.6337)**	7.23 (0.6113)
sw-my	17.84 (**0.7174**)	18.62 (0.7109)	**18.94** (0.7081)

*the values inside the parentheses are RIBES scores

Table 4: BLEU and RIBES scores of MWT and MSW pairs for PBSMT, HPBSMT and OSM using 1448 sentences

Evaluation Results

src-trg	MWT (word)-MSL(word)		
	PBSMT	***HPBSMT***	***OSM***
my-sl	**11.92** (0.7482)	11.30 **(0.7607)**	11.24 (0.7593)
sl-my	16.78 **(0.7913)**	16.04 (0.7898)	**17.25** (0.7842)

src-trg	MSL(syllable)-MWT(word)		
	PBSMT	***HPBSMT***	***OSM***
my-sl	14.12 (0.7674)	14.41 (0.7745)	**14.21 (0.7603)**
sl-my	20.21 (0.8107)	19.26 (0.8087)	**19.50 (0.8115)**

*the values inside the parentheses are RIBES scores

Table 5: BLEU and RIBES scores of MWT and MSL pairs for PBSMT, HPBSMT and OSM using 1448 sentences

Evaluation Results

src-trg	MSL(word)-MSW		
	PBSMT	***HPBSMT***	***OSM***
sl-sw	34.44 (0.8014)	34.99 (0.8049)	**37.54 (0.8280)**
sw-sl	52.66 (0.8754)	**52.79 (0.8756)**	49.99 (0.8675)

src-trg	MSW-MSL(syllable)		
	PBSMT	***HPBSMT***	***OSM***
sl-sw	33.99 (0.8206)	34.04 **(0.8260)**	**34.38** (0.8200)
sw-sl	49.47 (0.8660)	49.62 (0.8650)	**50.42 (0.8676)**

*the values inside the parentheses are RIBES scores

Table 6: BLEU and RIBES scores of MSL and MSW pairs for PBSMT, HPBSMT and OSM using 1448 sentences

Error Analysis

- Word Error Rate –WER
- It is dynamic programming to find an optimal alignment between the hypothesis of machine translation and the reference translation
- Formula of WER: $\text{WER}= \frac{S+I+D}{N}$
  - N = number of words
  - S = number of substituted words
  - I = number of inserted words
  - D = number of deleted words

Error Analysis

- Ref : ငါ အရေးပေါ် လူနာတင် ကား လိုချင် လို့ အမြန် ခေါ် ပါ ။
- Hyp : ငါ အရေးပေါ် လူနာတင် ကား အမြန် **** ခေါ်ပေး လိုချင် **** ။

WER errors:

Reference	Hypothesis	Error Type
လိုချင်	အမြန်	Substitution
လို့		Deletion
အမြန်	ခေါ်ပေး	Substitution
ခေါ်	လိုချင်	Substitution
ပါ		Deletion

- In this case, **S=3, D=2, C=5, N=10** for PBSMT and its **WER** is equal to **50%.**

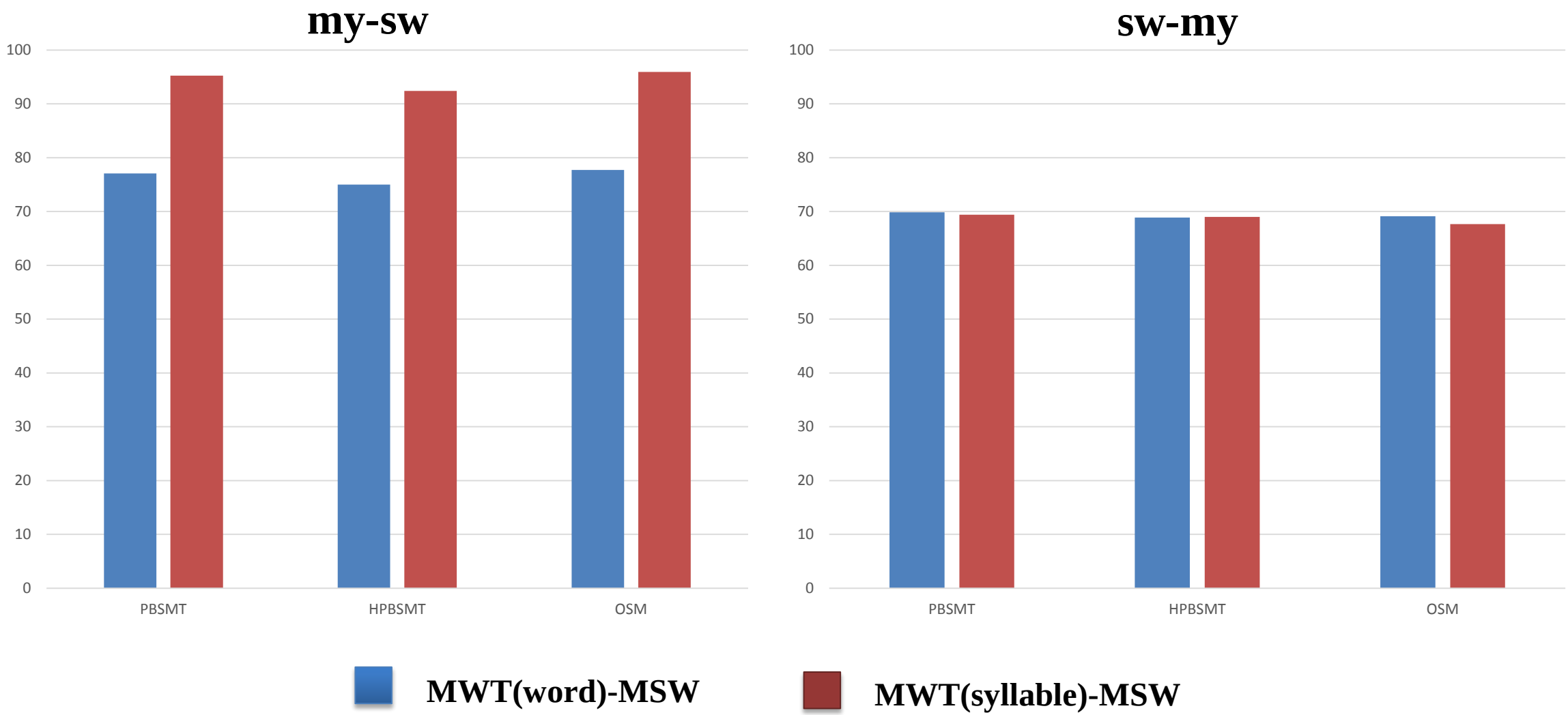


Figure 4 : WER of machine translation from Myanmar written text and Myanmar SignWriting using 888 sentences

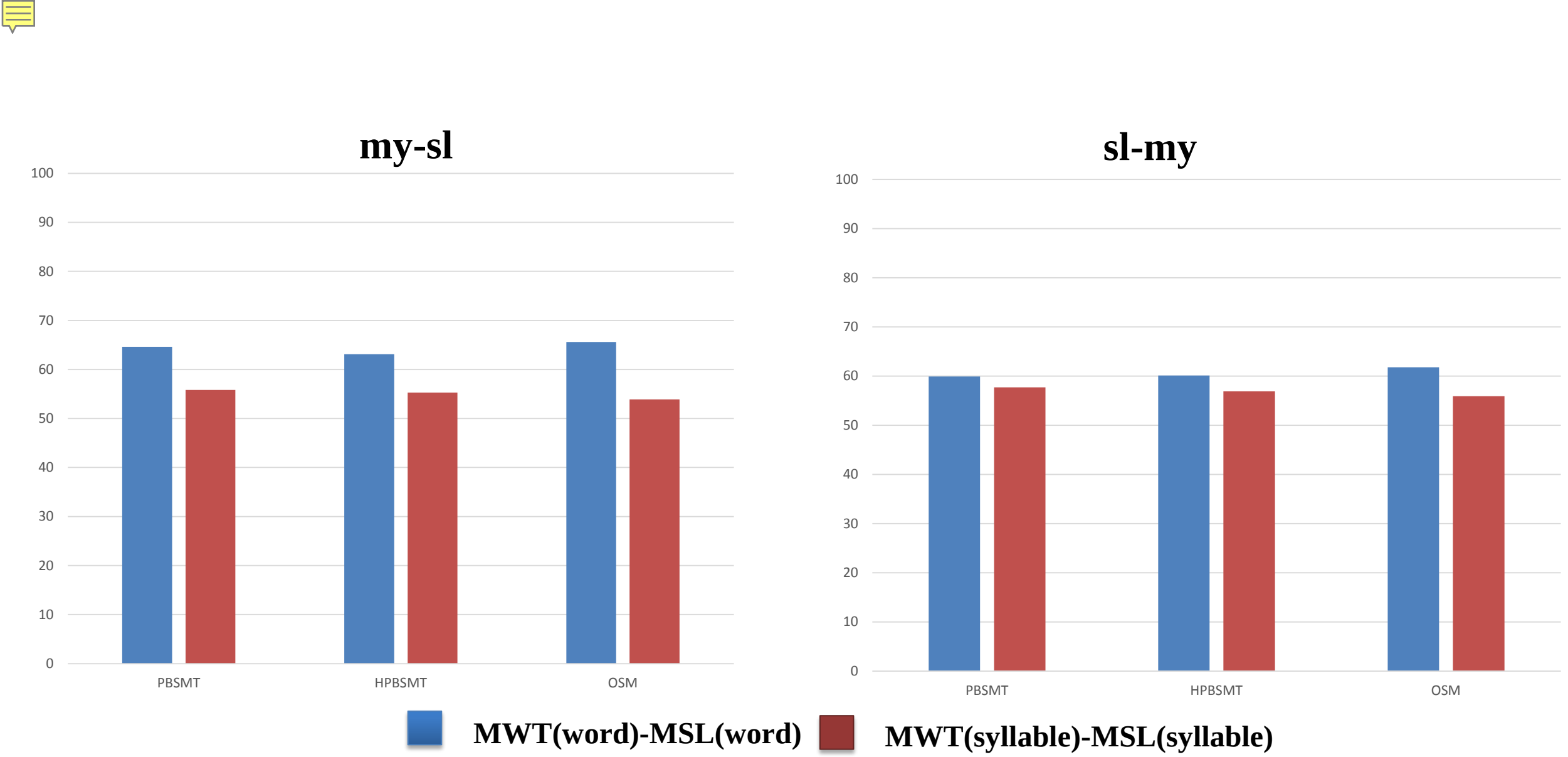


Figure 5 : WER of machine translation from Myanmar written text to Myanmar Sign Language using 888 sentences

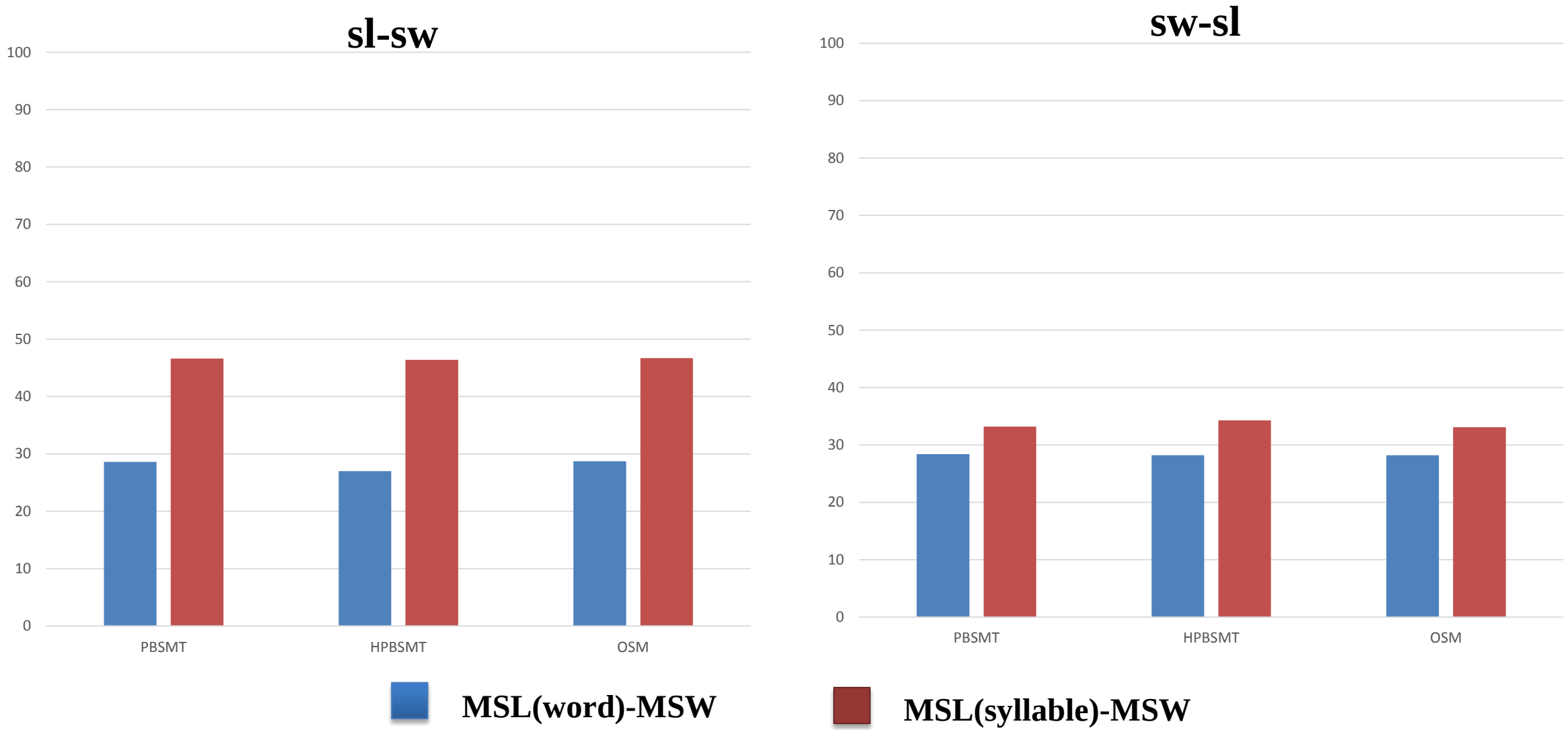


Figure 6 : WER of machine translation from Myanmar Sign Language to Myanmar SignWriting using 888 sentences

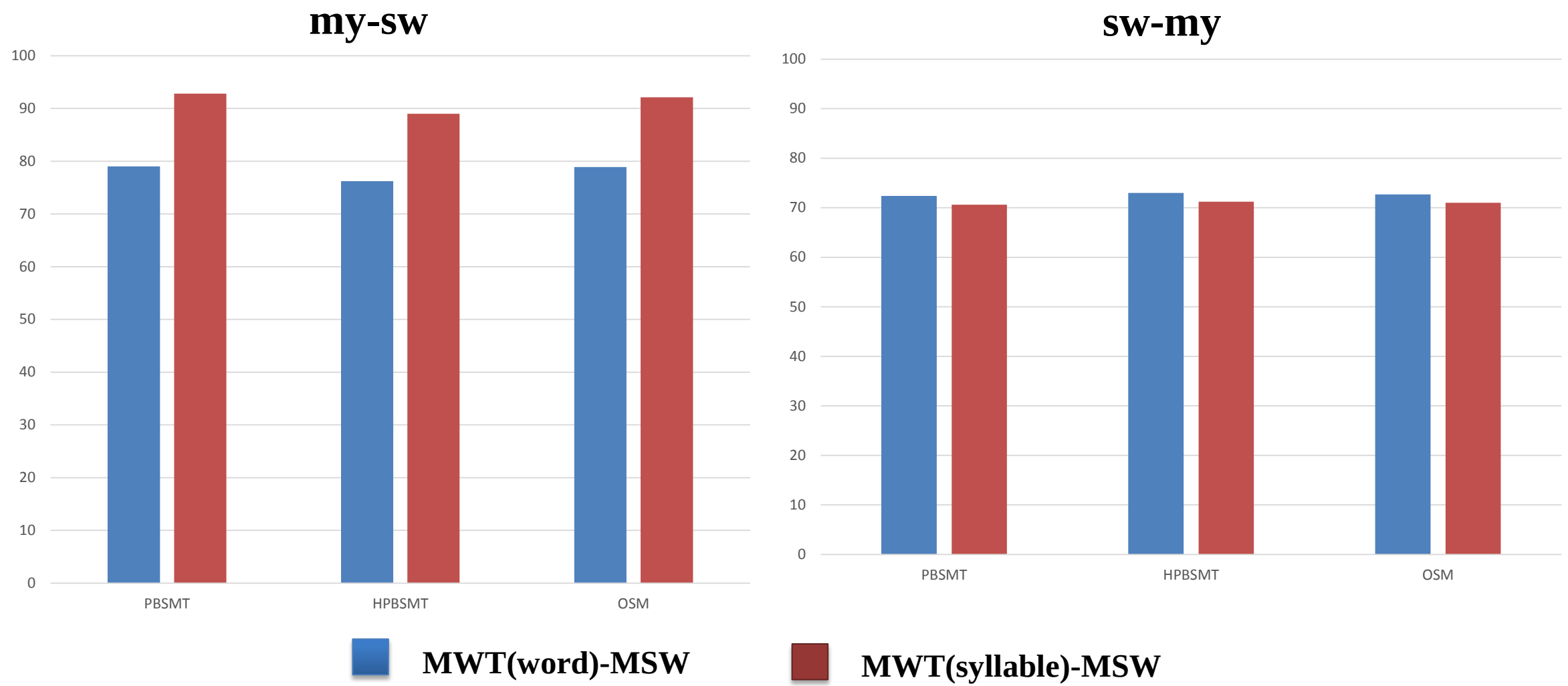


Figure 7 : WER of machine translation from Myanmar written text to Myanmar SignWriting using 1448 sentences

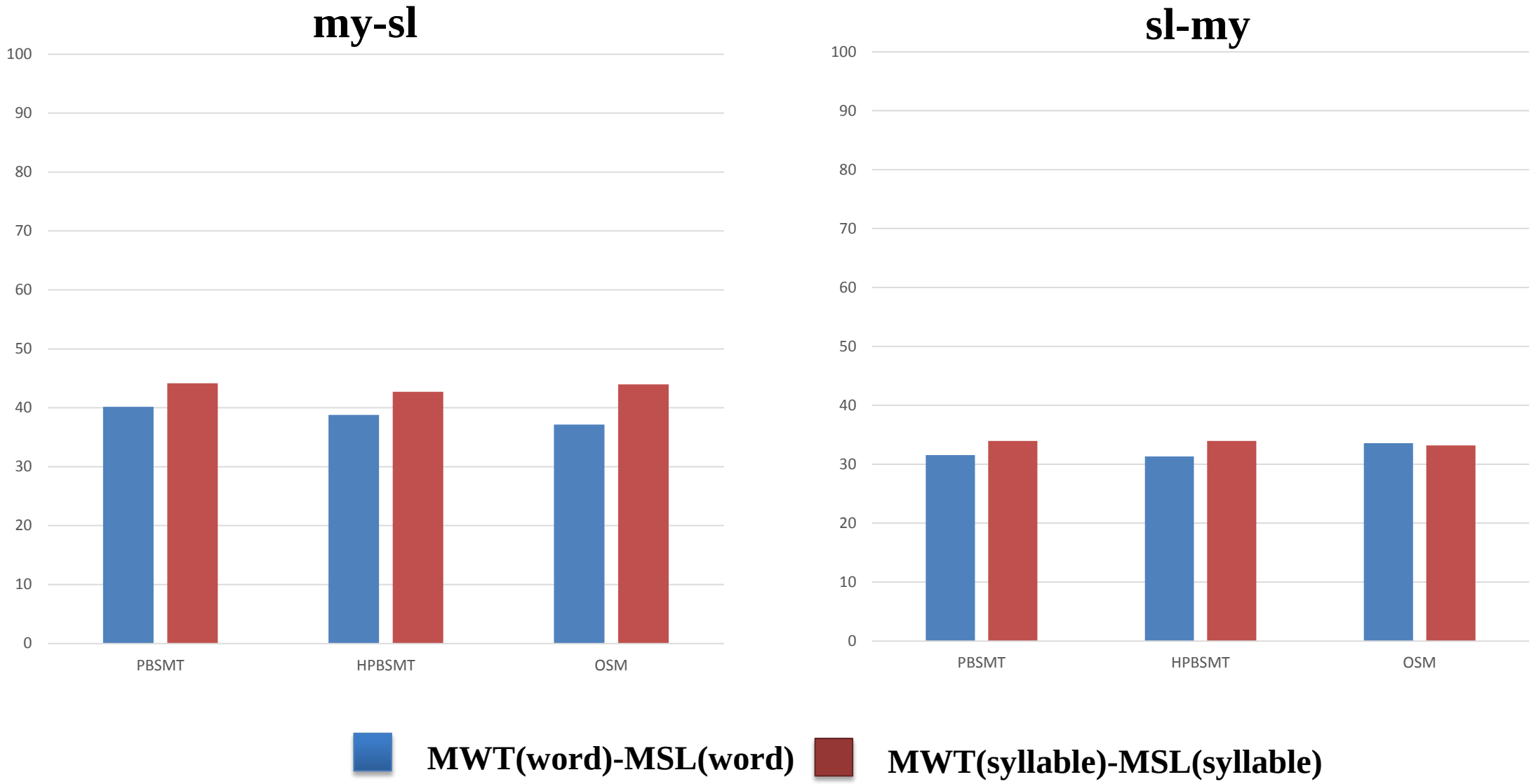


Figure 8 : WER of machine translation from Myanmar written text to Myanmar Sign Language using 1448 sentences

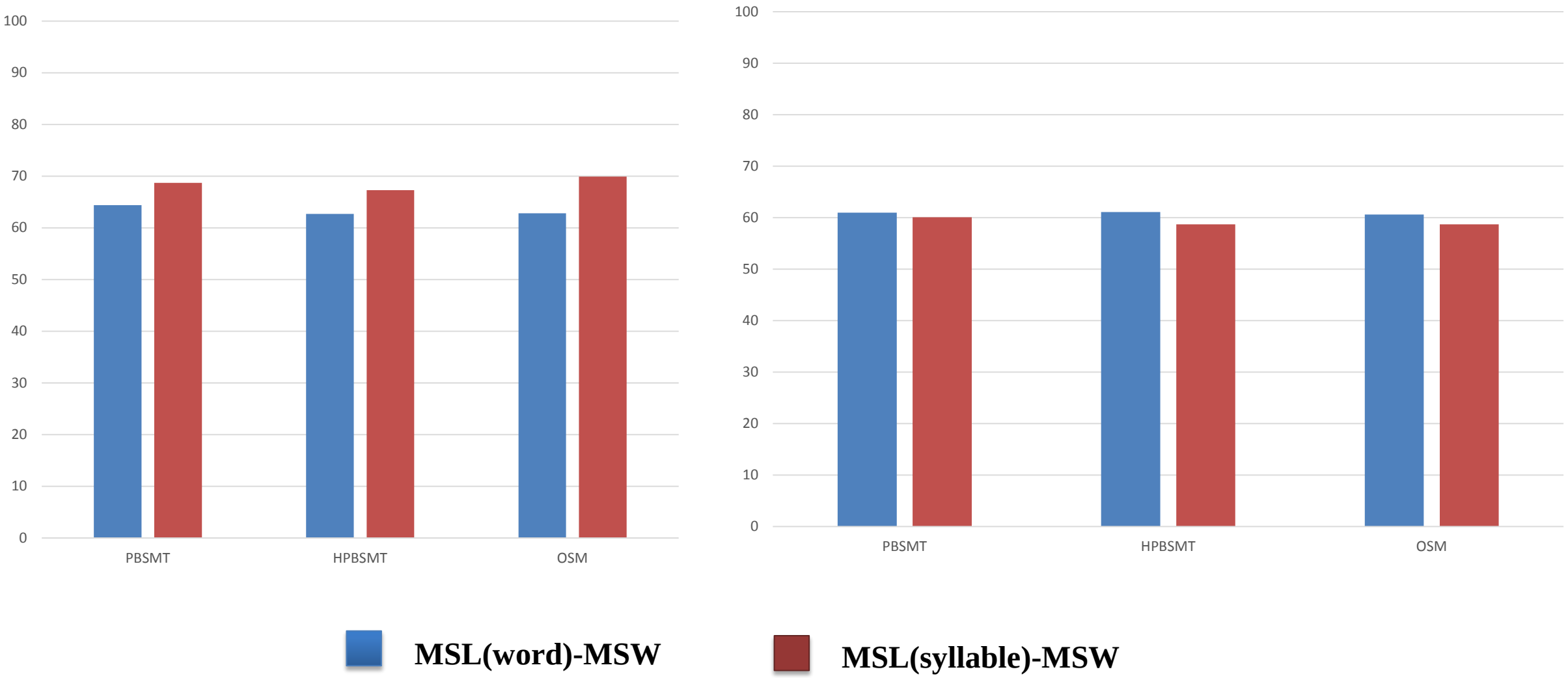


Figure 9 : WER of machine translation from Myanmar Sign Language to Myanmar SignWritingusing 1448 sentences

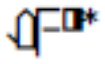
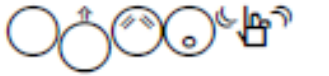
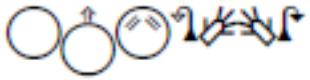
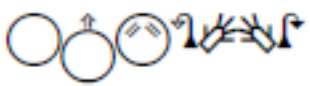
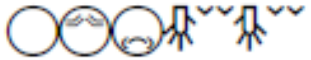
No	References	Hypothesis	Ref-Hyp Description in Myanmar Language	Freq
1			ဆရာဂန် -> ဆရာဂန်	3
2			ဘယ်လောက်လဲ -> ဘာလဲ	3
3				3
4			မသုံးနဲ့ -> မလုပ်နဲ့	3
5			မလုပ်နဲ့ -> မလုပ်နဲ့	3
6			ဘယ်သူ -> ဘာလဲ	2
7			ကယ်ဆယ် -> ကျေးဇူးပြု၍	2
8				2
9			ညစ်ပတ် -> ရေ	2
10			ရှိ -> မသေချာ	1

Table 7: Top 10 confusion pairs of OSM model in MWT-MSW translation

Error Analysis

- From detail analysis on confusion pairs of three SMT approaches, most of them are caused by four main reasons as follow:
  - **The nature of sign language,**
  - **Alignment error between source and target,**
  - **Some errors in the references or human mistakes,**
  - **Limited size of the training data**.

Conclusion

- The research introduced the first evaluation of machine translation between Myanmar Written Text, Myanmar Sign Language and Myanmar SignWriting using three SMT approaches.
- Translation performances are compared and error rate is analyzed.
- From the overall experimental results, OSM approach get the highest score in most of experiments.
- Word segmentation achieved better performance than syllable segmentation in both MWT and MSL.
- The system breaks down the language barrier and be able to get better communication between Deaf and hearing people.
- It helps to overcome the communication problem between Deaf and hearing people by introducing the SignWriting.

Future Works

- The future work will conduct experiments on SMT with SignWriting character level (i.e combination of basic symbol, filling symbol and spatial rotation symbol as a one SignWriting character) segmentation approach.
- The performance analysis can do with current SMT approaches and other machine translation approaches.

List of Publication

- Swe Zin Moe, Hlaing Myat Nwe, **Hnin Wai Wai Hlaing**, Ye Kyaw Thu, Hnin Aye Thant, Nandar Win Min, “Myanmar Sign Language (MSL) Corpus for Emergency Domain”, PACLING2017 conference, Yangon, Myanmar. (Demo and Poster)
- Swe Zin Moe, Ye Kyaw Thu, Hlaing Myat Nwe, **Hnin Wai Wai Hlaing**, Ni Htwe Aung, Hnin Aye Thant, Nandar Win Min, “Corpus Building for Machine Translation between Myanmar Sign Language and Myanmar Written Text”, World Deaf Day 2017, 14th Sept. 2017, Mandalay Community Center, Chan Aye Tharzan Township, Mandalay, Myanmar. (Poster)
- Swe Zin Moe, **Hnin Wai Wai Hlaing**, Ye Kyaw Thu, Hlaing Myat Nwe, Ni Htwe Aung, Hnin Aye Thant, Nandar Win Min, “□□□□□□ □□□□□□□□□□□□□□□□□□□□ □□ □□□□□□□□ □□□□□□”, International Day of Persons with Disabilities 2017, 3rd Dec. 2017, Wilson Hotel, No.31(E), Yangon-Mandalay Main Road, Maha Aung Myay Township, Mandalay, Myanmar. (Demo and Poster)
- Hlaing Myat Nwe, Ye Kyaw Thu, **Hnin Wai Wai Hlaing**, Swe Zin Moe, Ni Htwe Aung, Hnin Aye Thant, Nandar Win Min, "Two Fingerspelling Keyboard layouts for Myanmar SignWriting", International Day of Persons with Disabilities 2017, 3rd Dec. 2017, Wilson Hotel, No.31(E), Yangon-Mandalay Main Road, Maha Aung Myay Township, Mandalay, Myanmar. (Demo and Poster)

List of Publication

- Hlaing Myat Nwe, Ye Kyaw Thu, **Hnin Wai Wai Hlaing**, Swe Zin Moe, Ni Htwe Aung, Hnin Aye Thant, Nanda Win Min, "Two Fingerspelling Keyboard Layouts for Myanmar SignWriting", In Proceedings of ICCA2018, February 22-23, 2018, Yangon, Myanmar, pp. 290-298. (Paper)
- Swe Zin Moe, Ye Kyaw Thu, **Hnin Wai Wai Hlaing**, Hlaing Myat Nwe, Ni Htwe Aung, Hnin Aye Thant, Nandar Win Min, "Statistical Machine Translation between Myanmar Sign Language and Myanmar Written Text", In Proceedings of ICCA2018, February 22-23, 2018, Yangon, Myanmar, pp. 217-227. (Paper)
- **Hnin Wai Wai Hlaing**, Ye Kyaw Thu, Swe Zin Moe, Hlaing Myat Nwe, Ni Htwe Aung, Nandar Win Min, Hnin Aye Thant, “Statistical Machine Translation between Myanmar Sign Language and Myanmar SignWriting”, at the First IEEE International Symposium on Artificial Intelligence for ASEAN Development, ASEAN-AI2018, Phuket, Thailand, 26th March 2018. (Paper)

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
- https://github.com/ye-kyaw-thu/sylbreak
- https://github.com/ye-kyaw-thu/MSL4Emergency
- Wikipedia of Word Error Rate: https://en.wikipedia.org/wiki/Word_error_rate

Thank You for your kind attention.