The 1st ACIR+ WORKSHOP OF RESEARCH AND INNOVATION FOR STUDENTS (AWRIS 2024)
Theme: Towards innovative technology for paving the sustainable future in ASEAN

Statistical vs Neural Machine Translations for Khmer Braille

Authors:
Kimhuoy Yann, Ponleur Veng, Ye Kyaw Thu, Rottana Ly
Cambodia Academy of Digital Technology

July 20th, 2024

Table of contents

01

Introduction

Research Problem

- License Software Braille Translator
- Education document needed to support blind people

Research Objective

- **Research model that can translate from Khmer-to-Khmer Braille**
- **Implement this model to Website and App application**

02

Related Works

Ye Kyaw Thu et al Investigated SMT between Khmer and various languages using the Moses toolkit with different architectures (OSM, PBSMT, HPBSMT). https://aclanthology.org/Y15-1030
Zar Zar Hlaing et al Introduced Marian-NMT toolkit with different approach such as sequence-to-sequence (s2s) with both RNN-based. https://aclanthology.org/P18-4020
Zun Hlaing Moe et al Investigated statistical machine translation (SMT) between Myanmar text and Mu-Thit Braille, one of Myanmar's two Braille systems, utilizing IBM Models 1 and 2. https://aclanthology.org/2021.wat-1.6
Yu et al Propose a novel pre-training approach for Chinese–Braille translation, addressing the challenge of limited parallel data https://doi.org/10.1016/j.displa.2023.102506

03

Khmer Braille

- Invented by Cambodians and Thais in 1988 in the Saitou camp, Cambodian and in 1994, the Krousar Thmey community, the Ministry of Education, Youth, and Sport of Cambodia (MoEYS)and the Maryknoll Lay Missioners community improved Cambodian Braille to follow the structure of written Khmer.
- Use 6 dots per one cell system braille character.
- Due to complexity of Khmer character, there are 144 character to represent, and 6 dots system limited to 63 characters to expend the number 18 dots, or 3 braille cell was introduced.
- Position of Khmer braille really matter. Example:
  - ខ្មែរ = ខ + ◌្ + ម + រ
  - ខ្មែរ = ម + ខ + ◌្ + រ

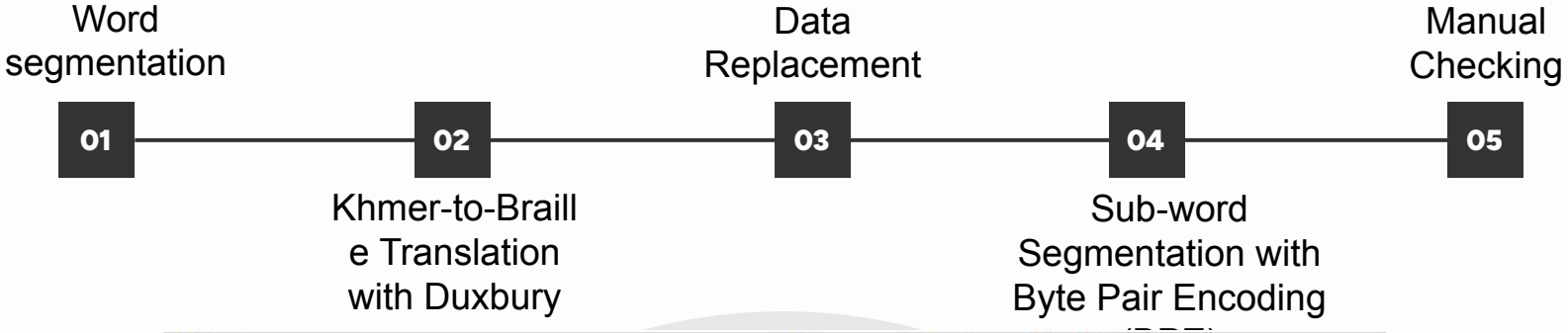
A-series	O-series
ក	គ
ខ	ឃ
ច	ជ
Special case	
ង	
ព	
ម	
រ	
វ	
ញ	
យ	
Consonant	Sub consonant
ក	◌្ក
ខ	◌្ខ
គ	◌្គ

Figure 1: Mappings for Khmer Braille consonants.

04

Khmer Braille Corpus Development

- 20K Khmer-Braille parallel corpus was developed which Khmer text was sourced from the Asian Language Treebank (ALT)and Braille, we utilized the Duxbury Braille Translator software.



Khmer text	Khmer-Braille (ASCII)
"ថ្វី ត្បិ តែ យើង សោក ស្តាយ ពី ការ បាត់ គាត់ គាត់ បន្សល់ ដំណែ លែ នឹង បង្កើន នូវ កំហឹង ដល់ សត្រូវ របស់ ជាតិ និង សាសនា ។"	0) vwe tvb / t LTt %, yRB s : g svt *, y & e g * r b * t 9 , g * t 9 , g * t 9 b , nvs , l 9 dyLTn , l , nLBRB b % RBvg , n , n 3 w gyhLBRB d , l 9 svrt 3 w rbs 9 , j * t / , n / RB s * s , n * = 0

Figure 2: Parallel sentence segmented with BPE sub-words

05

Experiment Setup

Statistical Machine Translation

- Phrase-Based Statistical Machine Translation (PBSMT)
- Operation Sequence Model (OSM)

Neural Machine Translatoin

- Sequence-to-sequence (Seq2Seq)
- Transformer

Corpus Statistics and Training

- Khmer-Braille corpus comprises a total of 20,106 parallel sentences and 715,013-word units.
- we allocated 18,088 parallel sentences for training, 1,000 for development, and 1,018 for testing.
- We train these SMT and NMT models using three AMD 3x RTX 3080ti 11GB GPUs, each paired with 64GB of RAM

06

Discussion

BiLingual Evaluation Understudy (BLEU Score)

Source-Target	Khmer-to-Braille	Braille-to-Khmer
PBSMT	85.61	70.64
OSM	85.65	70.63
Seq2Seq	72.55	65.41
Transformer	53.83	51.43

Table 1: BLEU Score result

Word Error Rate (WER)

Source-Target	Khmer-to-Braille	Braille-to-Khmer
PBSMT	13.30%	22.10%
OSM	**13.20%**	22.40%
Seq2Seq	25.60%	49.20%
Transformer	33.50%	55.90%

Table 2: Word Error Rate (WER) result

07

Conclusion

- First-time study results of SMT and NMT for Khmer text to Braille translation and vice versa
- The development of a Khmer-Braille parallel corpus, which we plan to release publicly after further checking and data cleaning
- The experimental results, in terms of BLEU and WER scores, show that SMT significantly achieved better translation performance than NMT
- We plan to extend the Braille corpus to ASEAN languages.
- Further study especially on NMT and current research trends of fine-tuning with large language model approaches

Thank You
For
Your Listening

VKU
NHÂN BẢN – PHỤNG SỰ – KHAI PHÓNG

ICT FISU
FACULTIES · INSTITUTES · SCHOOLS · UNIVERSITIES

ACIR

CITA2024