Statistical Machine Translation between Myanmar (Burmese) and Rakhine (Arakanese)

Thazin Myint Oo[†], Ye Kyaw Thu[∫, λ], Khin Mar Soe[†]

[†]Natural Language Processing Lab., *University of Computer Studies, Yangon, Myanmar*

[∫]*Artificial Intelligence Lab., Okayama Prefectural University (OPU), Japan*

[λ]*Language and Speech Science Research Lab., Waseda University, Japan*

thazinmyintoo@ucsy.edu.mm, ye@c.oka-pu.ac.jp, khinmarsoe@ucsy.edu.mm

Sections

- Introduction
- Related Work
- Rakhine Language
- Methodology
- Experiments
- Results and Discussion
- Evaluation
- Error Anslysis
- Conclusion

Introduction

- main motivation for this research is to investigate SMT performance for Myanmar (Burmese) and Rakhine (Arakanese) language pair

- five major machine translation approaches applied to low-resource languages. PBSMT, HPBSMT, tree-to-string (T2S), string-to-tree (S2T) and OSM translation methods
- there is no publicly available tree parser for Rakhine language
- cannot apply S2T and T2S approaches for Myanmar-Rakhine language pair
- the machine translation experiments were carried out using PBSMT, HPBSMT and OSM

Related work

- "Machine Translation Experiments on PADIC: A Parallel Arabic DIalect Corpus" , Karima Meftouh et al. 2015
- built PADIC (Parallel Arabic Dialect Corpus) corpus from scratch
- experiments on **cross dialect Arabic machine translation**
- PADIC is composed of dialects from both the Maghreb and the Middle-East
- **interesting results** were achieved even with the limited corpora of 6,400 parallel sentences

Related Work

- "A Hybrid Approach to Statistical Machine Translation Between Standard and Dialectal Varieties" , Friedrich Neubarth et al. , 2013
- suffers from **data sparsity**
- combining **word-level and character-level models** can yield good results even with small training data
- by exploiting the relative proximity between the two varieties ,arising with the translation between standard Austrian German and Viennese dialect
- used **hybrid approach** of rule-based preprocessing and PBSMT for getting better performance

Related work

- "Machine Translation of Low-Resource Spoken Dialects: Strategies for Normalizing Swiss German" Pierre-Edouard Honnet et al. ,2017
- **proposed solutions** for the machine translation of a family of dialects, Swiss German
- **three strategies** for normalizing Swiss German input in order to address the regional and spelling diversity
- The results show that **character-based neural MT** was the most promising one for text normalization and that in combination with **PBSMT achieved 36% BLEU score**

Rakhine(Arakanese)

Rakhine Language

- Rakhine language used the script Arakanese or Rakkhawanna Akkhara before at least the 8th century A.D. current Rakhine script is exactly the same with Myanmar script
- Arakanese language notably retains on /r/ sound (i.e. “ရ”) that has become /j/ sound (i.e. “ယ”) in Burmese. And thus “ ကြား:” (“to hear” in English) and “ကျား:” (“tiger” in English) pronounced differently as “kya” and “kra” in Rakhine language.
- three main dialects corresponding to the five administrative districts of Rakhine division
- The total population for all countries is nearly about 3,000,000

Arakanese language notably retains on /r/ sound (i.e. “ရ”) that has become /j/ sound (i.e. “ယ”) in Burmese. And thus “ ကြား:” (“to hear” in English) and “ကျား:” (“tiger” in English) pronounced differently as “kya” and “kra" in Rakhine language

Examples

my: လုံချည် တစ် ထည် ဘယ်လောက်လဲ ။

rk: ဒယော တစ် ထည် ဇာလောက်လေး ။

("How much for a longyi?" in English)

my: အဘွား နေ မကောင်းဘူး ။

rk: အဘောင်သျှင် နီ မကောင်းပါ ။

("Grandma is not feeling well." in English)

my: ကလေး များ ကစား နေကြတယ် ။

rk: အချေ တိ ကဇော် နီကတ်တေ ။

("Children are playing." in English)

Methodology

- Phrase-Based Statistical Machine Translation
- is based on **phrasal units**
- model typically gives better translation performance than word-based models
- 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

$$\boldsymbol{argmax_e P(e|f)} = \boldsymbol{argmax_e P(f|e) P(e)}$$

Methodology

Hierarchical Phrase-Based Statistical Machine Translation

- a model based on **synchronous context-free grammar**
- The model is able to be learned from a corpus of **unannotated parallel text**
- The **advantage** offers over the phrase-based approach is that the hierarchical structure is able to represent the word re-ordering process.
- particularly applicable to language pairs that require **long-distance re-ordering** during the translation process

[X][X] ကနေ [X] ||| [X][X] ကနိုန် [X]
[X][X] ကနေ [X] ||| [X][X] ကနီ [X]
[X][X] ကနေ [X] ||| [X][X] ကနေ [X]
[X][X] ကနေ [X] ||| [X][X] ကပင်ဆိုကေ [X]
[X][X] ကနေ [X] ||| [X][X] ဂနီ [X]

Methodology

Operation Sequence Model

- 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
- bilingual language model that also integrates reordering information
- 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

Source: Please sit here
Target: ကျေးဇူးပြုပြီး ဒီမှာ ထိုင်
Operation 1: Generate(Please, ကျေးဇူးပြုပြီး)
Operation 2: Insert Gap
Operation 3: Generate (here, ဒီမှာ)
Operation 4: Jump Back (1)
Operation 5: Generate (sit, ထိုင်)

Experiments

Corpus Statistics

- used 18,373 Myanmar sentences (without name entity tags) of the ASEAN-MT Parallel Corpus
- Word segmentation for Rakhine was done manually and there are exactly 123,018 words in total
- 10-fold cross-validation experiments and used 14,023 to 14,078 sentences for training, 2,475 to 2,485 sentences for development and 1,810 to 1,875 sentences for evaluation respectively

Word Segmentation

root word and suffixe(s) are separated such as "စား ဗျာယ်", "စား ပီးဗျာယ်", "စား ဖို့ဗျာယ်". Here, "စား" ("eat" in English)

Rakhine adverb words such as "အဂယောင့်" ("really" in English), "အမြန်" ("quickly" in English) are also considered as one word

Rakhine word "ကလိန့်မရေျ တိ" (ladies) is segmented as two words "ကလိန့်မရေျ" and the particle "တိ"

Word Segmentation

Rakhine compound word "ဖေ့သာ + အိတ်" ("money" + "bag" in English) is written as one word "ဖေ့သာအိတ်" ("wallet" in English)

Rakhine word "အကြိုစေ့နှစ်ခတ်" ("two coins" in English) is segmented as "အကြိုစေ့ နှစ် ခတ်"

Moses SMT System

- Moses toolkit 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++
- The alignment was symmetrize by grow-diag-final and heuristic and lexicalized reordering model was trained with the msd-bidirectional-fe option
- We use KenLM or training the 5-gram language model with modified Kneser-Ney discounting
- Minimum error rate training (MERT) was used to tune the decoder parameters and the decoding was done using the Moses decoder (version 2.1.1)

Evaluation

Bilingual Evaluation Understudy (BLEU)

Rank-based Intuitive Bilingual Evaluation Measure (RIBES)

Results and Discussion

- **Table 1. Average BLEU and RIBES scores for PBSMT, HPBSMT and OSM**

- OSM method give highest BLEU and RIBES score
- Rk-my machine translation is better(around 3 BLEU and 0.02 RIBES score higher)

Table 1. Average BLEU and RIBES scores for PBSMT, HPBSMT and OSM

src-tgt	PBSMT	HPBSMT	OSM
my-rk	57.68 (0.9077)	57.70 (0.9073)	**57.88 (0.9085)**
rk-my	60.58 (0.9233)	60.42 (0.9230)	**60.86 (0.9239)**

Result and Discussion

- two more 10-folds cross-validation experiments (for PBSMT and OSM) with and without reordering
- both PBSMT and OSM gave approximately the same results
- only local reordering is enough for the Myanmar-Rakhine language pair

Table 2. Average BLEU and RIBES scores for PBSMT and OSM with reordering and without reordering

src-trg	PBSMT	OSM
my-rk (without)	57.70 (0.9078)	57.89 (0.9086)
my-rk (reordering)	57.69 (0.9077)	57.89 (0.9086)
rk-my (without)	60.56 (0.9232)	60.86 (0.9239)
rk-my (reordering)	60.57 (0.9232)	60.86 (0.9239)

Error Analysis

- SCLITE scoring method for calculating the erroneous words in WER
- first make an alignment of the hypothesis (the translated sentences) and the reference
- perform a global minimization of the Levenshtein distance function which weights the cost of correct words, insertions (I), deletions (D), substitutions (S) and the number of words in the reference (N).
- The formula for WER can be stated as Equation

$$WER = \frac{(N_i+N_d+N_s)\times 100}{N_d+N_s+N_c}$$

Scores: (#C #S #D #I) 2 1 0 1
REF : *** ဈာအိမ်မှာ မင်းနိုလေး ။
HYP : ဈာ အိမ်မှာ မင်းနိုလေး ။
Eval : I S

S=1, D=0, I=1, C=1, N=2

WER is equal to 66.67%.

Error Analysis

Table 3. Average WER% for PBSMT, HPBSMT and OSM with nearly 1,800 sentences test data (lower is better)

src-tgt	PBSMT	HPBSMT	OSM
my-rk	25.89%	25.94%	**25.78%**
rk-my	22.46%	22.53%	**22.26%**

- WER% for all three approaches are very closed to each other.
- OSM achieved the lowest WER%
- HPSMT method is highest WER%.

Manual Analysis

- WER calculation does not consider the contextual and syntactic roles of a word.
- some extra words are containing in the translated outputs of all three SMT approaches especially for Myanmar to Rakhine machine translation.

SOURCE:
နောက် တချက်ချေမာပင် မျောက်တိ က အလားတူ
လိုက်လုပ် ကတ်ရေ ။

REFERENCE:
နောက် ခဏချင်းမှာပဲ မျောက်တိ က အလားတူ
လိုက်လုပ် ကတ်ရေ ။

HYP of PBSMT:
နောက် ခဏချင်းမှာပဲ မျောက်တိ က အလားတူ က
လိုက်လုပ် ကတ်ရေ ။

Table 4. The top 10 confusion pairs of PBSMT model for Myanmar- Rakhine

Freq	Confusion Pair (REF→HYP)
15	ပါ ။ ==> ။
13	ငါ ==> ကျွန်တော်
12	ရှိ ==> သူရှိ
12	အကျွန် ==> ကျွန်တော်
10	ကို ==> ယင်းချင့်ကို
10	လား ==> ပါလား
9	နှန့် ==> နှန့်
9	လိမ့်မေ ==> လိမ့်မယ်
9	လေး။ ==> ။
8	ကတ်တော ==> ကတ်ရော

Conclusion

- contributes the first PBSMT, HPBSMT and OSM machine translation evaluations from Myanmar to Rakhine and Rakhine to Myanmar.
- 18K Myanmar-Rakhine parallel corpus that we constructed to analyze the behavior of a dialectal Myanmar-Rakhine machine translation.
- higher BLEU and RIBES scores can be achieved for Rakhine-Myanmar language pair even with the limited data.
- detail analysis on confusion pairs of machine translation between Myanmar-Rakhine and Rakhine-Myanmar.
- In the future we plan to test PBSMT, HPBSMT and OSM models with other Myanmar dialect languages such as Yaw and Dawei.