NLP Class Final Project Presentation

Myanmar Part of Speech Tagging Based on Machine Translation

Phyo Thu Htet, Naing Linn Phyo, Thiha Nyein

28.1.2021 (Thu)

Outline

1. Abstract
2. Introduction
3. Data
4. Data Preprocessing
5. Methodology
6. Error Analysis
7. Results and Discussion
8. Conclusion
9. References
10. Team
11. Task Assignment

Abstract

- This paper compares the performances achieved by Phrase-Based Statistical Machine Translation system(PBSMT), Neural Machine Translation based on sequence model of LSTM (Long Short Term Memory), and Attention-Based Neural Machine Translation systems (NMT) when translating Part-Of-Speech tagging (POS tagging) in Myanmar Language.
- Part of speeches are very fundamental in a language and various languages had done in tagging part-of-speech and the use of tag sets vary upon each own language.
- Important in Linguistics and part of speech tags can be used in building of Parse-tree(used in NERs) and for lemmatizers and many other NLP tasks.
- The systems are built on a parallel Myanmar-POS tagging corpus with word segmented 10k text.
- Through the research, PBSMT outperforms NMT. PBSMT also has a much higher evaluation score in comparison.
- Experiment's performances are evaluated with RIBES score, BLEU score, and ChrF++.

Introduction

- POS tagging is the activity of marking up a word in a text (corpus) as corresponding to a particular part of speech.
- As it has the relationship with adjacent and related words in a phrase, sentence, or paragraph, Part-Of-Speech Tagging stands as a fundamental role in NLP research processes an important port in Linguistics.
- Myanmar Language also needs to be POS tagged(Grammatical tagging).
- Like other translations (eg: english to spanish), Myanmar and POS tags can use the machine translation approach.
- Myanmar sentence is input as source language and POS would be the target language. E.g.

  Myanmar: ဦးသန့် အား ၎င်း မွေးဖွား ရာ ပန်းတနော် မြို့ ကို ရည်စူး ၍ ပန်းတနော် ဦးသန့် ဟု အမည်တွင် ခဲ့ သည် ။

  POS Tag: n ppm pron v part n n ppm v conj n n part v part ppm punc

Cont'd

- Neural Machine Translation, especially Attention-based models, and Phrase-based Machine Translation both recently become the choices among translate methods.
- In this paper, we focused on both **Encoder Decoder Model** and also **Encoder Decoder** with Attention mechanism.
- Statistical methods have normally based on many approaches but choose to perform with a phrase-based mechanism.
- Phrase tables mapped one-to-one in parallel. The data texts are word-segmented.
- In this paper there are 15 part-of-speech tags for Myanmar Language.

Data

- Datas are from GitHub of Dr.Ye Kyaw Thu.
- Data are segmented and Myanmar-POS tags parallel corpus.
- Maximum words per a sentence is cut to be 150 and left 10k in data sentences.
- Sample of the data can be described as

  ပထမ မှာ သားရေ ပေါ် တွင် ကပ် ၍ ပေါက် သော အမွေး နု များ ဖြစ် သည် ။<|||>adj ppm n n ppm v conj v part n adj part v ppm punc

  ခရစ်နှစ် ၁၈၈၆ ခုနှစ် တွင် ဘင်္ဂလား ခြေလျင်တပ် မှ ဗိုလ်မှူးကြီး မေ ကို အစွဲပြု ၍ မေမြို့ ဟု ဖြိတိသျှ တို့ က ခေါ်တွင် ခဲ့ သည် ။<|||>n num n ppm n n ppm n n ppm v conj n part n part ppm v part ppm punc

Cont’d

1. abb as Abbreviation (E.g. အ.မ.က, အ.လ.က),
2. adj as Adjective (E.g. ဝသော, ပိန်သော),
3. adv as Adverb (E.g.နှေးနှေး),
4. conj as Conjunction (E.g. နှင့်, ဖြင့်),
5. fw as Foreign Word (E.g. 1, 2, Facebook),
6. int as Interjection (E.g. အောင်မလေး),
7. n as Noun (E.g. ကား, နွား),
8. num as Number (E.g. ၁, ၂, ၃),
9. part as Particle (E.g. ပြီး, များ),
10. ppm as Post-positional Marker (E.g. သည်, က, ကို , သို့, မှာ, တွ င်),

Cont’d

11. pron as Pronoun(E.g. ကျွန်မ, သင်, သူ , သူ မ),
12. punc as Punctuation(E.g. ။, ၊),
13. sb as Symbol(E.g. %, $),
14. tn as Text Number (E.g. တစ်, နှစ်)
15. v as Verb (E.g. လုပ်ရှား)

Data Preprocessing

- Data Formatting is performed.

  Original Data: ယ့ခု/n လ/n တွင်/ppm ပျားရည်/n နှင့်/conj ပျားဖယောင်း/n များ/part ကို/ppm စုဆောင်း/v ကြ/part သည်/ppm ဟု/part ခန့်မှန်း/v နိုင်/part သည်/ppm ။/punc

  Formatted Data: ယခု လ တွင် ပျားရည် နှင့် ပျားဖယောင်း များ ကို စုဆောင်း ကြ သည် ဟု ခန့်မှန်း နိုင် သည် ။<|||>n n ppm n conj n part ppm v part ppm part v part ppm punc
- Data is shuffled and divided into training, development, and testing.
- In neural machine translation, start” and “end” are introduced in the target sentences.

  E.g. start_ n ppm n n ppm ppm n adj n adj v part part ppm punc _end

Cont’d

- Formatting to fit into pandas dataframe is executed.
- The preprocessing stages of word indexing, sentence indexing, and padding are performed in LSTM, and LSTM with attention.
- Perl, Python, and shell programs are used.
- Mosesdecoder and

Cont'd

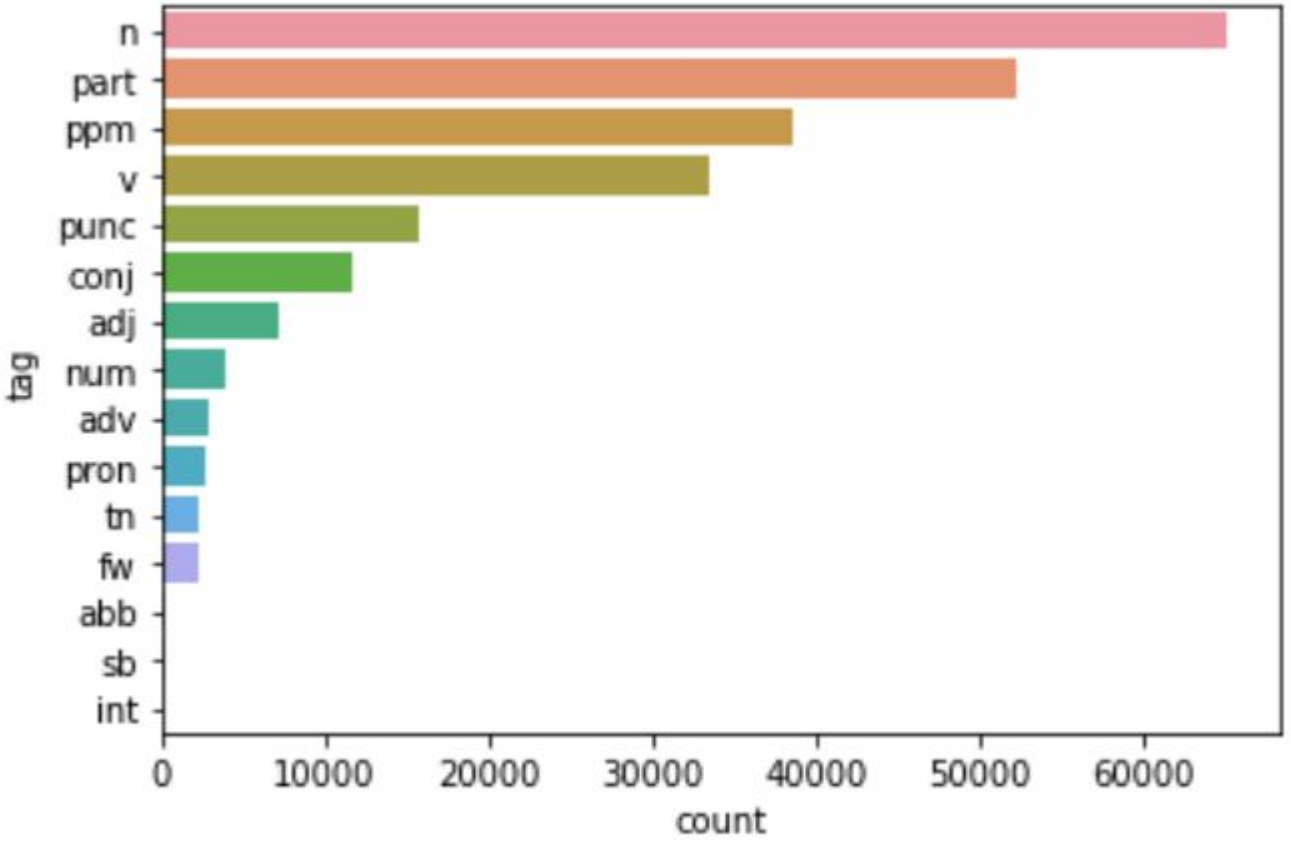


FIGURE I: Analysis of the number of tags

Methodology

- Statistical Machine Translation
  - Phrase Based Statistical Machine Translation(PBSMT)
  - MosesDecoder are used to implement PBSMT
- Neural machine Translation
  - Encoder -Decoder
  - Encoder -Decoder model with Attention Mechanism
  - TensorFlow are used to implement NMT

Phrase Based Statistical Machine Translation (PBSMT)

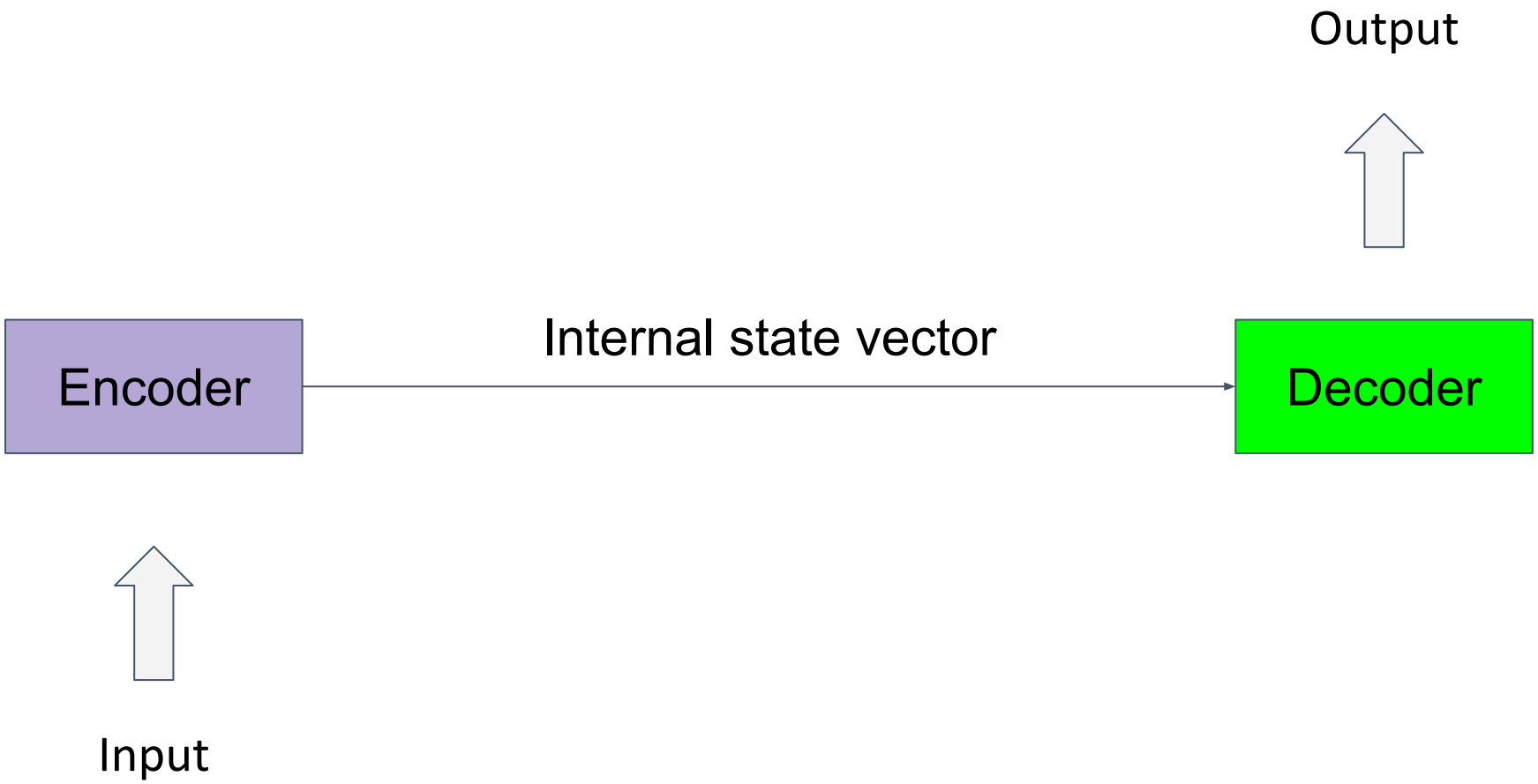
- A PBSMT translation model is based on phrasal units
- A phrase is simply a contiguous sequence of words and generally, not a linguistically motivated phrase
- A phrase-based translation model typically gives better translation performance than word-based models
- Phrase-based translation model consisting of
  - phrase-pair probabilities extracted from corpus
  - a basic reordering model
  - an algorithm to extract the phrases to build a phrase-table
- To find best translation **ê** that maximizes the translation probability **P(f)** given the source sentences

Encoder-Decoder Model

- A way of using RNN for sequence-to-sequence prediction problems.
- Basically LSTM (or) GRU [Note: *LSTM is used in our experiment*]

  //Involves two recurrent neural networks,
  - //Encoder to encode the input sequence
  - //Decoder to decode the encoded input sequence into the target sequence
- Usage of encoder-decoder
  - Chatbots
  - Machine Translation
  - Text summary
  - Image captioning

Architecture

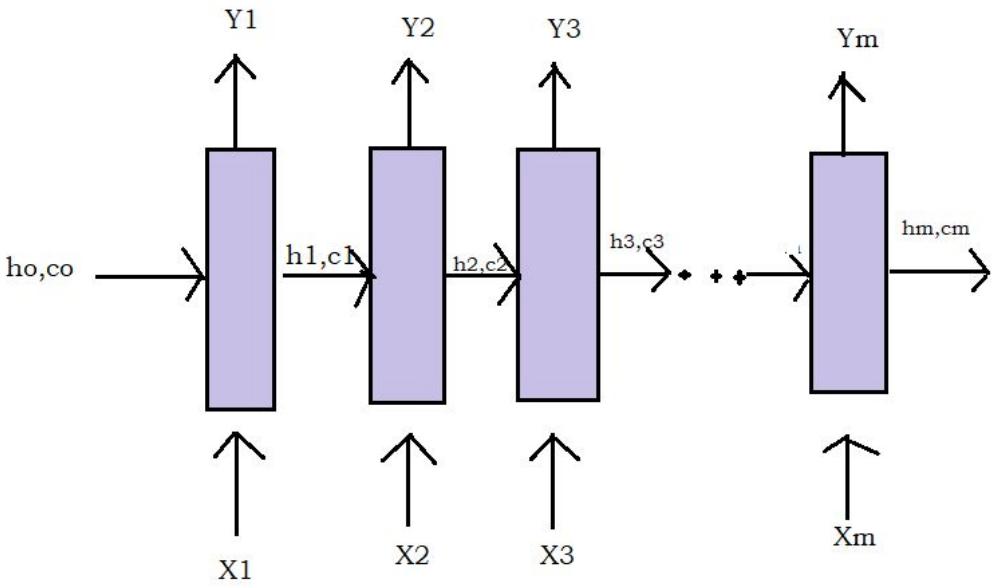


Cont’d

- Encoder
  - accepts a single element of the input sequence at each time step,
  - process it,
  - collects information for that element and propagates it forward.
- Intermediate vector
  - **final internal state** produced from the encoder path of model
  - contains information about the entire sequence to help the decoder make accurate predictions
- Decoder
  - give the entire sentence
  - it predicts an output at each time step

Encoder

- **Takes** the input sequence and **encapsulates** the information as the internal state vectors
- Outputs of the encoder are rejected and only internal states are used



LSTM for encoder

Cont’d

LSTM takes only one element at a time, so if the input sequence is of length **m**, then LSTM takes m time steps to read the entire sequence.

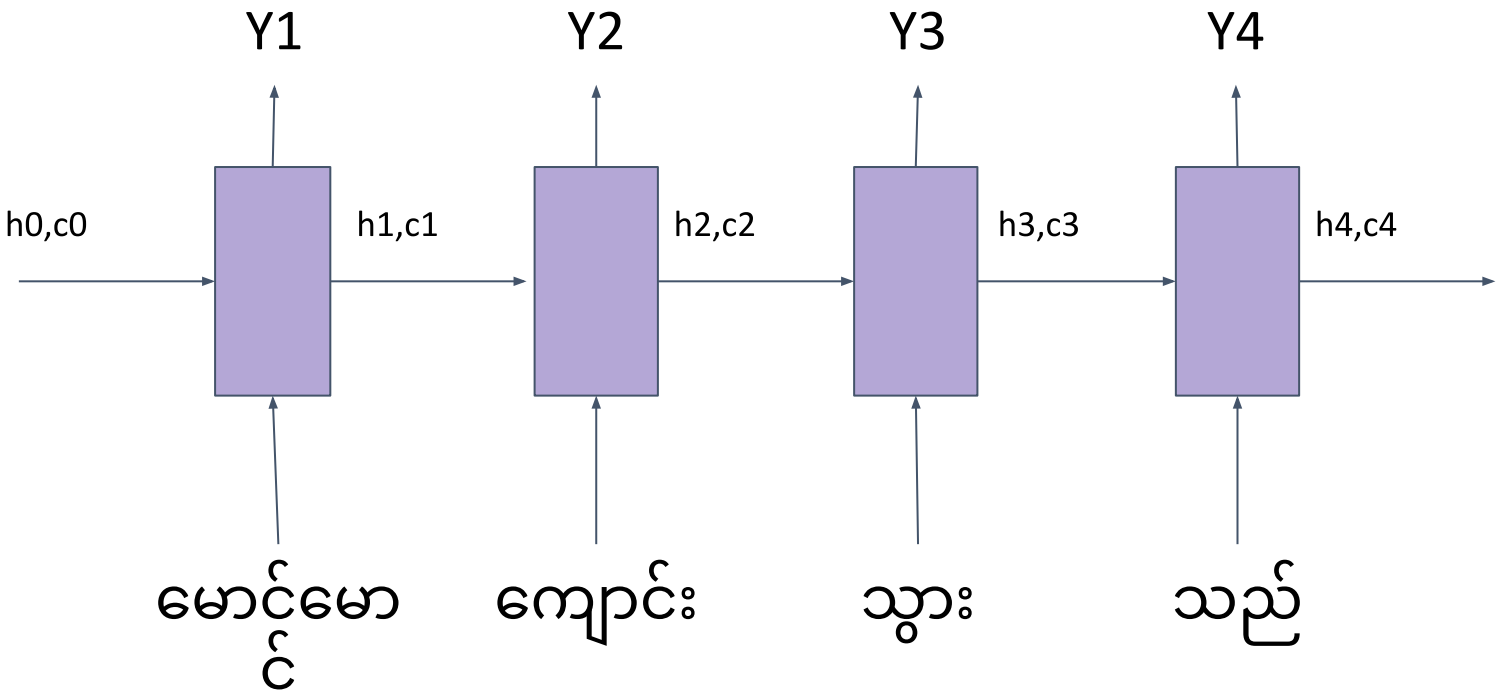
- **Xt** is the input at time step **t**.
- **ht** and **ct** are internal states at time step **t** of the LSTM and for GRU there is only one internal state ht.
- **Yt** is the output at time step **t**

Example

မောင်မောင် ကျောင်း သွား သည် <|||> n n v ppm

X1 = မောင်မောင် X2 = ကျောင်း

X3 = သွား X4 = သည်

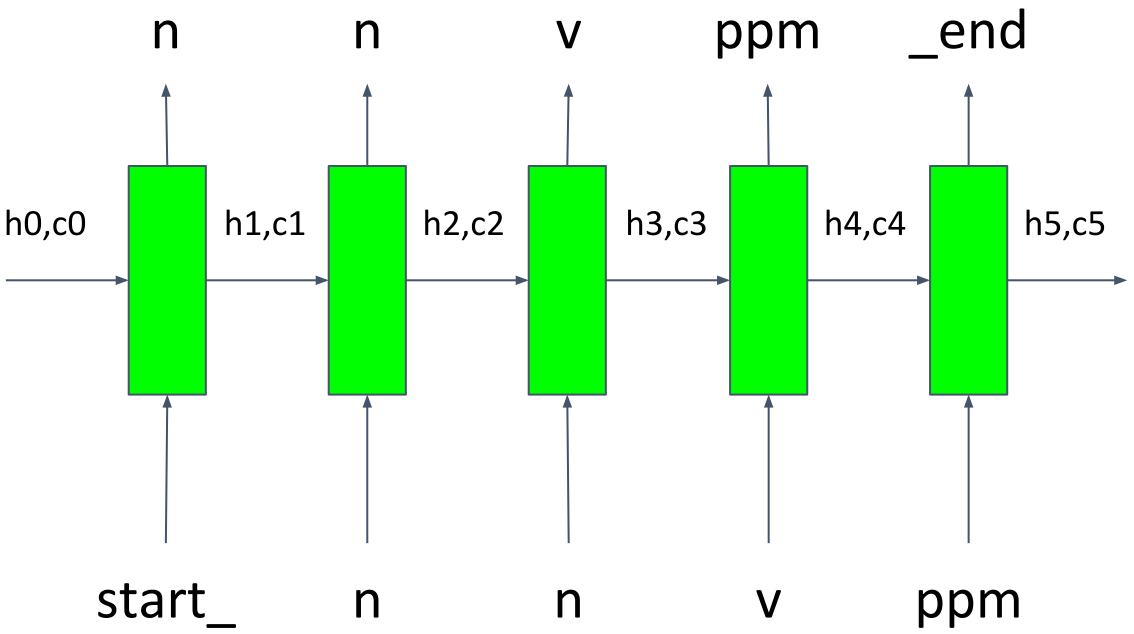


Cont’d

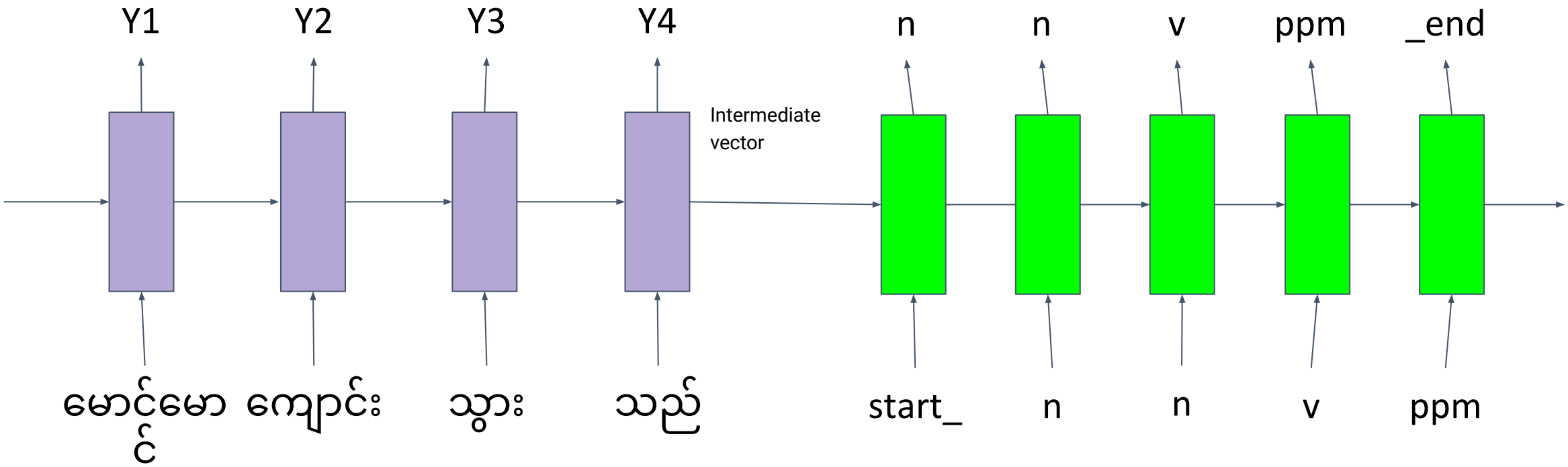
- The outputs of each time step **Y1,Y2,Y3,Y4** are
  - not used in Encoder-Decoder Model
  - Used in Encoder-Decoder with Attention Mechanism to feed to the attention layer

Decoder

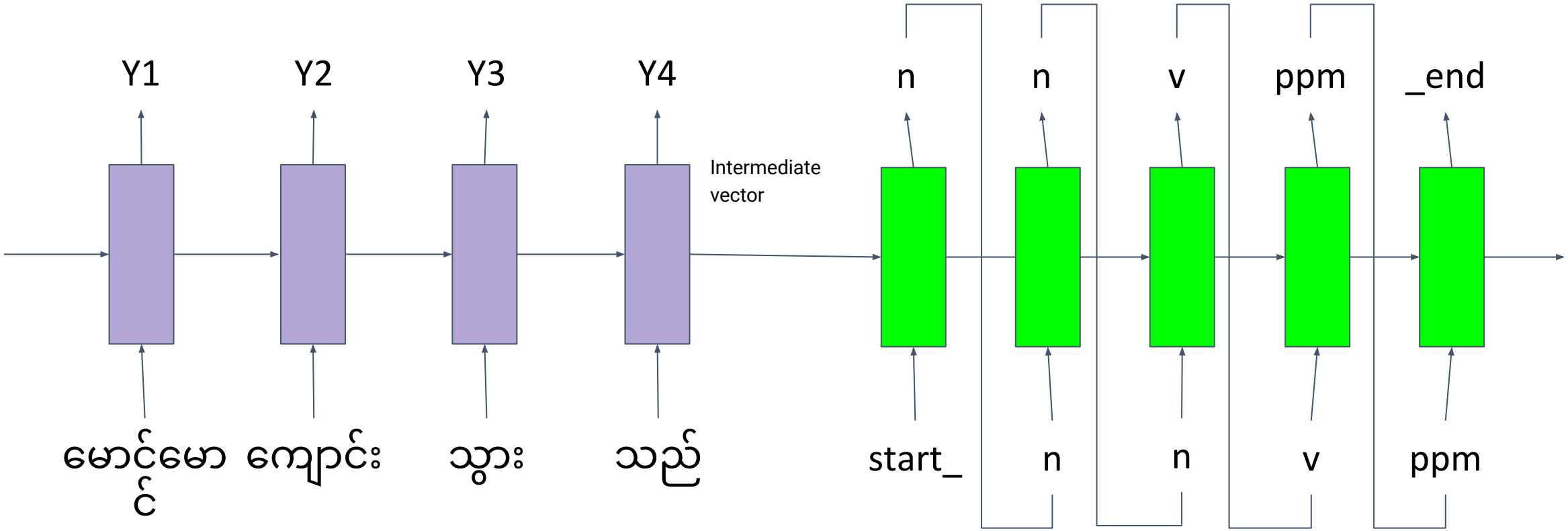
- The working of the decoder is different during the training and testing
- To recognize the the starting and end of the sequence,
  - Add “start_ ” at the beginning of the output sequence
  - “ _end” at the end of the output sequence



Encoder-Decoder Training Time



Encoder Decoder Testing Time



Attention Mechanism

- **It is difficult for the encoder model to memorize long sequences and convert it into a fixed-length vector**
- it predicts the next word by concentrating on a few relevant parts of the sequence rather than looking on the entire sequence.

Two Types of Attention Mechanism

- Global Attention
  - Global Attention is those attention in which all the hidden state vectors of the encoder are passed to get the context vector.
- Local Attention
  - Local Attention is those attention in which only a few hidden state vectors of encoder are considered for the generation of context vectors.

In this research, global attention is used.

Result and Discussion

	BLEU	RIBES	ChrF++ (c6+w2-avgF)
PBSMT	**0.7727**	**0.9726**	**89.05**
LSTM	0.4529	0.8647	75.25
LSTM with attention	0.7699	0.9659	85.57

TABLE I BLEU, RIBES, and chrF ++ scores for PBSMT, LSTM, LSTM with attention of Myanmar POS Tagging (Bold numbers indicated the highest scores)

Cont'd

- The performance of PBSMT in all of the categories of BLEU, RIBES, and ChrF ++ is the best.

  BLEU score (0.7727), RIBES score (0.9726), and ChrF ++ (89.05)
- On the other hand, the lowest score results in LSTM Model.

  BLEU score (0.4529), RIBES score (0.8646), and ChrF ++ (75.25)
- LSTM enhanced with attention Mechanism is also working well in the evaluation scores.

  BLEU score (0.7699), RIBES (0.9659), and ChrF++(85.57)
- For the current data, the architecture of LSTM can not perform well regarding BLEU Score (0.4529).

Error Analysis - PBSMT

Input: မြန်မာ့ တေးဂီတ ကို ပတ်ဝိုင်း ၊ ကြေးဝိုင်း ၊ ပတ္တလား နှင့် လေမှုတ်တူရိယာ များ ဖြစ်သည့် နှဲ ၊ နှဲကြီး ၊ ပလေ ၊ ဝါးလက်ခုပ် နှင့် ကြုံးတပ်တူရိယာ များ အား ဆိုင်းဝိုင်း ခေါ် သံစုံတီးဝိုင်း ဖွဲ့ ၍ တီးခတ် ကြ သည် ။

Reference: n n ppm n punc n punc n conj n part v part n punc n punc n punc n conj n part ppm n v n v conj v part ppm punc

Hypothesis: n n ppm **ပတ်ဝိုင်း** punc **ကြေးဝိုင်း** punc **ပတ္တလား** conj **လေမှုတ်တူရိယာ** part v part v punc **နဲ့ကြီး** punc **ပလေ** punc **ဝါးလက်ခုပ်** conj **ကြုံးတပ်တူရိယာ** part ppm **ဆိုင်းဝိုင်း** v **သံစုံတီးဝိုင်း** v conj **တီးခတ်** part ppm punc

Cont'd

- Input: ၁၉၉၂ ခုနှစ် တွင် ရှန်ကုန် မြို့ စမ်းချောင်း မြို့နယ် မှ ဒေါ်နော်မာ နှင့် အိမ်ထောင်ကျ ကာ ဒေါ် ဦးဦးမေ ၊ ဦးကျော်ကျော် ၊ ဒေါက်တာ အောင်သော် ၊ ဒေါမော်မော် နှင့် ဒေါသန်းသန်းဆွေ စသည့် သားသမီး များ ထွန်းကား ခဲ့ သည် ။

  Reference: num n ppm n n n n ppm n ppm v conj n punc n punc n n punc n conj n part n part v part ppm punc

  Hypothesis: num n ppm n n n n ppm ppm v conj **ဒေါ်နော်မာ ဒေါ် ဦးဦးမေ** punc **ဦးကျော်ကျော်** punc n **အောင်သော်** punc **ဒေါ်မော်မော်** conj **ဒေါသန်းသန်းဆွေ** part n part v part ppm punc

- Most of the errors are Names, Numbers that not include in data, Foreign words and single words.

Error Analysis - Sequence to Sequence model

Confusion Pair	INSERTIONS	DELETIONS
1: 169 -> part ==> n	1: 561 -> part	1: 670 -> part
2: 168 -> v ==> n	2: 553 -> n	2: 601 -> n
3: 128 -> n ==> part	3: 348 -> v	3: 466 -> v
4: 107 -> n ==> v	4: 295 -> ppm	4: 428 -> ppm
5: 95 -> v ==> part	5: 176 -> conj	5: 260 -> conj
6: 86 -> part ==> v	6: 137 -> adj	6: 201 -> adj
7: 80 -> ppm ==> n	7: 71 -> punc	7: 114 -> punc
8: 71 -> ppm ==> part	8: 66 -> adv	8: 104 -> adv
9: 67 -> n ==> ppm	9: 43 -> num	9: 69 -> num
10: 63 -> part ==> ppm	10: 41 -> fw	10: 63 -> fw

Error Analysis - Sequence to Sequence with attention

- Confusion-pair: (('v', 'part'), 86), (('v', 'n'), 84), (('part', 'v'), 82), (('n', 'v'), 69), (('part', 'n'), 64)
- Most of the OOV are predicted as “n”.

Original Input: ၁၈၆၀ ခုနှစ် တွင် ဒီလရှယ်လီဘရားသားစ် က ခရစ်ယာန် သာသနာပြု ကျောင်း များ ကို တည်ဆောက် ခဲ့ ကြ သည် ။

Input: ၁၈၆၀ ခုနှစ် တွင် OOV က ခရစ်ယာန် သာသနာပြု ကျောင်း များ ကို တည်ဆောက် ခဲ့ ကြ သည် ။

Reference: num n ppm n ppm n **v** n part ppm v part part ppm punc

Hypothesis: num n ppm n ppm n **n** n part ppm v part part ppm punc

Cont’d

- Since POS tagging is based on the approach of Machine Translation, the predicted length is not the same as original length of the sentence. (The maximum length of predicted POS sentences is 25)

  Input: သို့နှင့် မဂ္ဂဇင်း မှ တစ်ဆင့် သတင်းစာ ကို ပါ တိုးချဲ့ လိုက် သောအခါ တွင် ဘက်ပတစ် ကျောင်း သို့ မ ပြန် တော့ ဘဲ ထို မဂ္ဂဇင်း ၊ သတင်းစာ နှစ် ခု စလုံး တွင် ပင် တည်းဖြတ် သည့် ဘက် မှ ဆက်လက် လုပ်ကိုင် လေ တော့ သည် ။

  Reference: conj n ppm adv n ppm part v part conj ppm n n ppm part v part part adj n punc n tn part part ppm part v part n ppm adv v part part ppm punc (Length: 37)
  Hypothesis: conj n ppm adv n ppm part v part conj ppm n n ppm part v part part adj n punc n tn part part (Length: 25)

Conclusion

- According to the results, Myanmar POS Tagging based on Machine Translation is also useful by giving the appropriate performance.
- Phrase-Based Statistical Machine Translation provides better results than the other approaches used in this research.
- Neural Machine Translation with the attention mechanism can also provide excellent results (96.33% in RIBES score).
- With more data, the results can be different.

Future Work

- The process of hyperparameter tuning will also be performed.
- The research will also extend by using the metrics like Accuracy, Precision, Recall, and so-on.
- More data will be introduced to the current data.
- Supplementing with the segmentation schemes like character-level segmentation, syllable level segmentation, and sub-word level segmentation of a sentence would be interesting.

References

[1] José Carlos Rosales Núñez, ”A Comparison between NMT and PBSMT Performance for Translating Noisy User-Generated Content”, Université Paris Sud, LIMSI.

[2] Guillem Gasc´o i Mora and Joan Andreu S´anchez Peir´o, ”Part-of-Speech Tagging Based on Machine Translation Techniques”, Departament de Sistemes Inform‘atics i Computacif´o Universitat Polit‘encia de Val‘encia Cam´ı de Vera s/n, 46022 Val‘encia (Spain) .

[3] Xiaocheng Feng, ”Enhanced Neural Machine Translation by Joint Decoding with Word and POS-tagging Sequences”, Springer Science+Business Media, LLC, part of Springer Nature 2020.

[4] Koehn, Philipp, and Och, Franz Josef and Marcu, Daniel, “Stastatistical phrase-based translation,” Proceedings of the 2003 Con- ference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology Volume 1, 2003, pp. 48–54.

[5] Lucia Specia„ “Tutorial, Fundamental and New Approachesto Statistical Machine Translation,” International Conference.

Cont'd

[6] D. Kauchak and R. Barzilay, "Paraphrasing for automatic evaluation," Asso. Com. Ling. New York City, USA, vol. Pro. Hum.Lang. Tech. Conf. NAACL, Main Conference, pp. 455–462, June 2006.

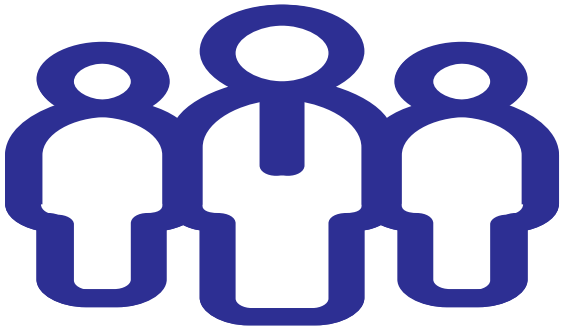
[7] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico,N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin and E. Herbst, "Moses: Open source toolkit for statistical machine translation," Asso. Com. Linguistics, booktitle. Proc. ACL-08: HLT, Proc. 45th Ann. Meet. Asso.Com. Ling. Comp. vol. Proc. Demo and Poster Sessions, Prague, Czech Republic, pp. 177–180, June 2007.

[8] F. Braune, A. Gojun and A. Fraser, "Long-distance reordering during search for hierarchical phrase-based SMT," In Proc. of the 16th Ann. Conf. of the Euro. Asso. for Mac. Translation, Trento, Italy, pp. 177–184. 2012.

[9] Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. J. (2002).

[10] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, Hajime Tsukada Automatic Evaluation of Translation Quality for Distant Language Pairs

Team



Team Name - Goldilock

Team Leader - Phyo Thu Htet

Other Members - Naing Linn Phyo, Thiha Nyein

Working Session - 1 hour (5:30pm -6:30pm or 8pm-9pm) of every day in general

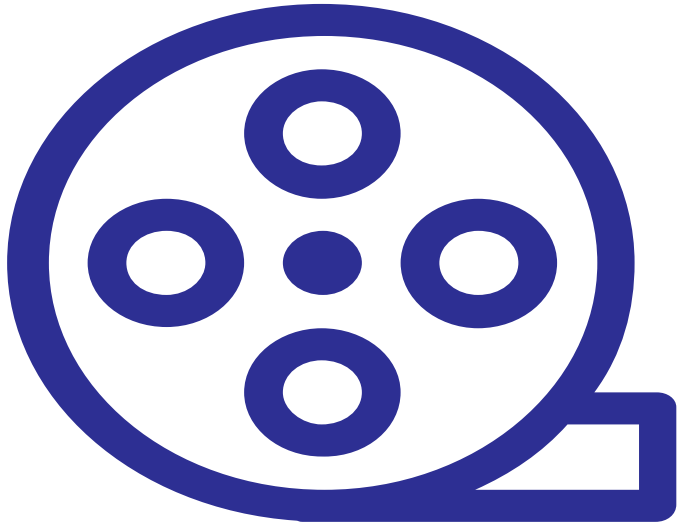
Channel - Messenger, Microsoft Team

Task Assignment

Member	Machine Translation	Paper	Other
Naing Linn Phyo	PBSMT	Abstract, Introduction, Related Works, POS Tags	Power Point (As of Paper)
Thiha Nyein	LSTM (Encode-Decoder)	Methodology, Evaluation (Subsection of Experiement Setup)	Power Point (As of Paper)
Phyo Thu Htet	LSTM with Attention	Result and Discussion, Experiment Setup (Except Evaluation), Conclusion, Latex	Power Point (As of Paper), Meeting Setup and Management


UNIVERSITY OF TECHNOLOGY
UTYCC (EST-2010)
YATANARPON CYBER CITY

Thank You



Note

This powerpoint template is created by The Hsu Kyaw and owned by Phyo Thu Htet.