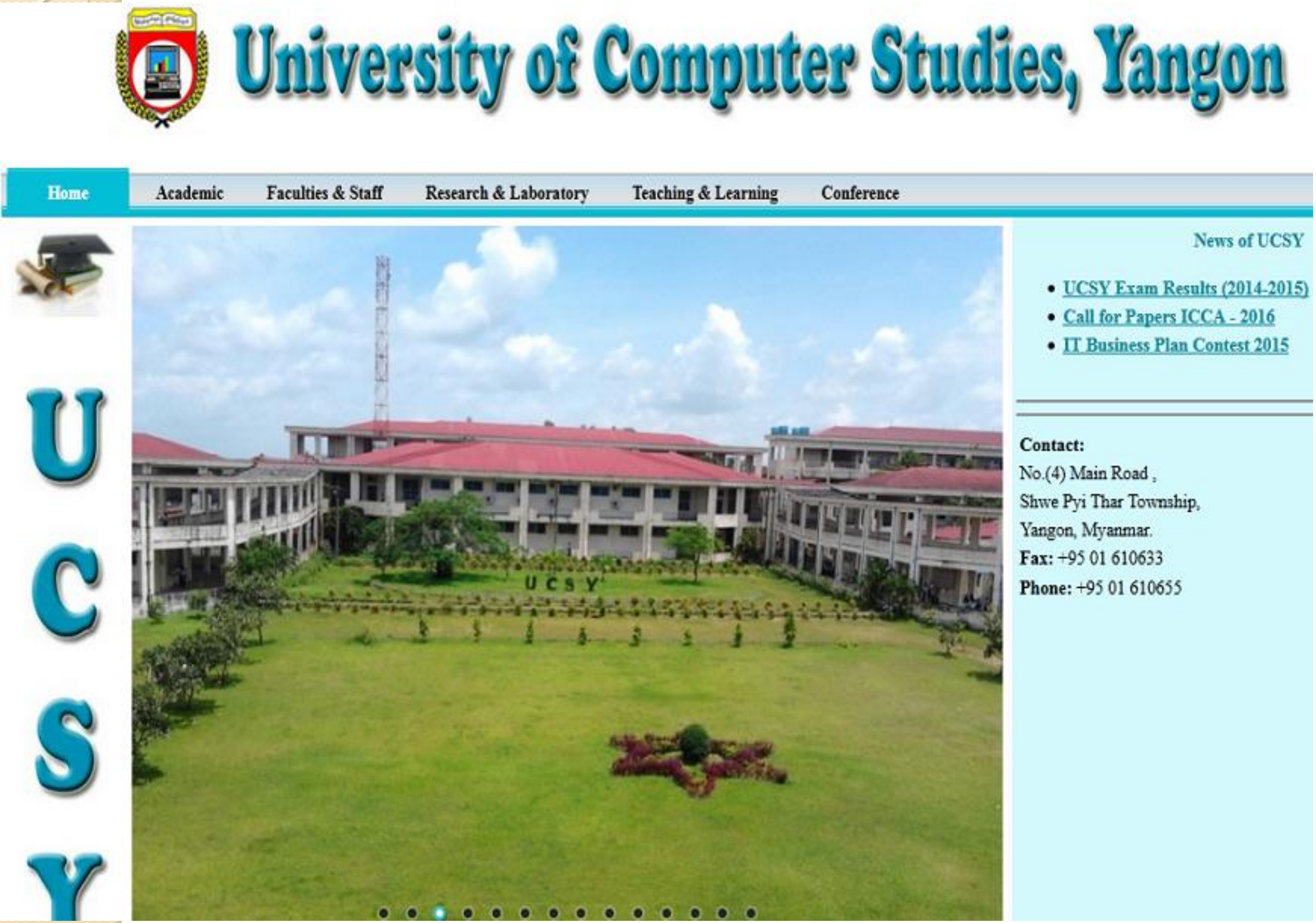
Myanmar NLP research and Usefulness of ALT data

Dr. Khin Mar Soe
Professor
NLP Lab, UCSY

26-11-2015

Contents

- ❖ Introduction to UCSY
- ❖ Introduction to UCSY NLP Lab
- ❖ Current Myanmar NLP Research
- ❖ Usefulness of ALT Data
- ❖ Conclusion



University of Computer Studies, Yangon

News of UCSY

- UCSY Exam Results (2014-2015)
- Call for Papers ICCA - 2016
- IT Business Plan Contest 2015

Contact:
No.(4) Main Road ,
Shwe Pyi Thar Township,
Yangon, Myanmar.
Fax: +95 01 610633
Phone: +95 01 610655

http://www.ucsy.edu.mm

Natural Language Processing Lab in UCSY

- started in 2006 at University of Computer Studies, Yangon (UCSY) under Ministry of Science and Technology.
- Some of the works of the NLP lab are available online:
  - Network-based ASEAN Languages Translation Public Service (http://www.aseanmt.org)
  - English to Myanmar Statistical Machine Translation System (http://www.nlpresearch-ucsy.edu.mm/NLP_UCSY/mtapplication.html)
  - Myanmar-English-Myanmar bilingual dictionary (http://www.nlpresearch-ucsy.edu.mm/NLP_UCSY/dictionaryapplication.html)
  - Myanmar Word Segmentation (http://www.nlpresearch-ucsy.edu.mm/NLP_UCSY/wsandpos.html)

Research Collaboration

- NECTEC (Thailand National Electronics and Computer Technology Center)
- NICT (National Institute of Information and Communication Technology)

- For the purpose of
  - joint researches/projects,
  - researcher exchange,
  - publishing conference papers, journals and articles,
  - doing joint NLP workshops.

NLP Lab

NLP Lab Members

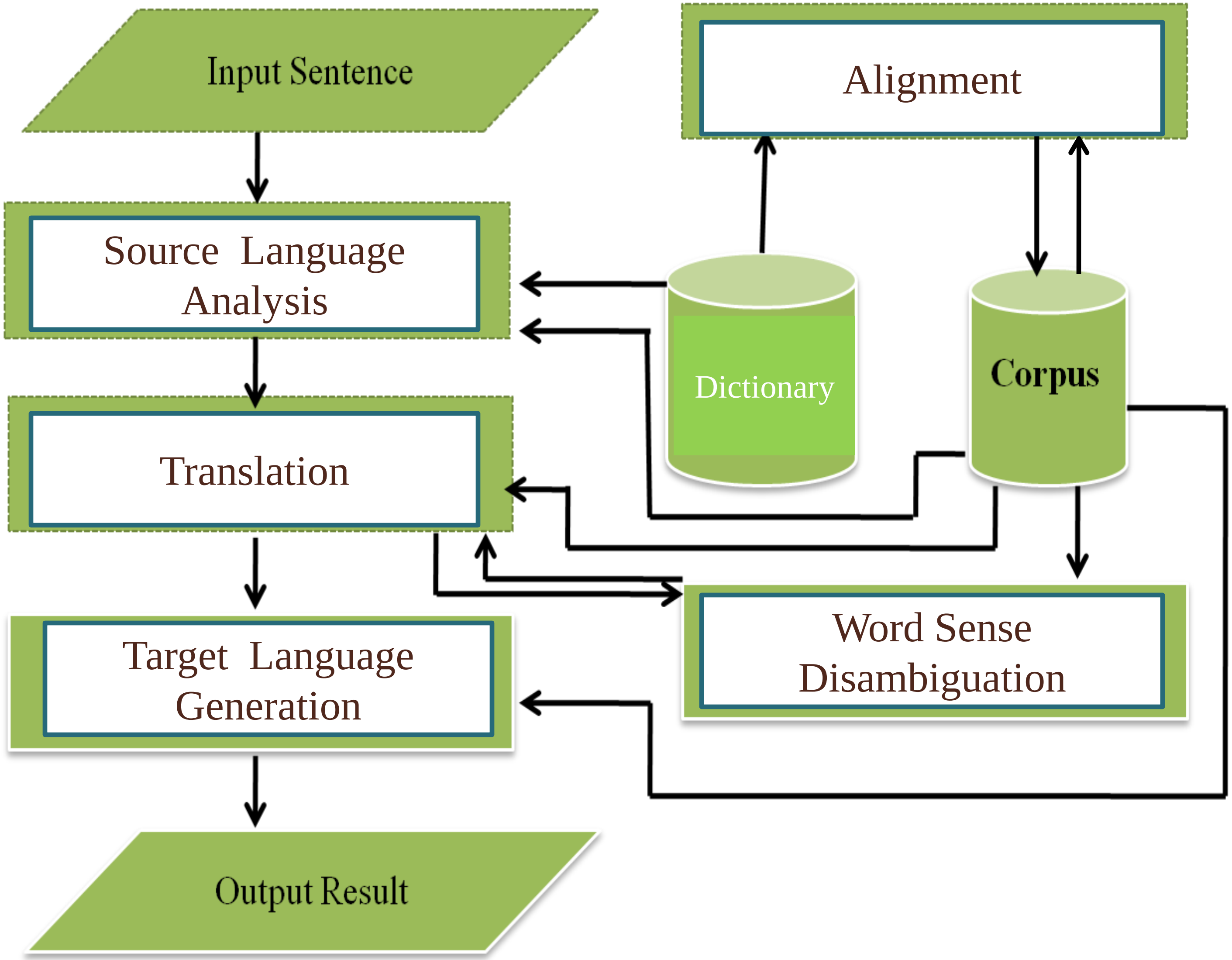
NLP Research

Aim of Research

- to overcome language barrier
- to be applied conveniently in systems that are used by Myanmar

- **Domain of Research**
  - Myanmar-English-Myanmar Machine Translation
  - Automatic Speech Recognition
  - Text to Speech
  - Myanmar Information Retrieval
  - Myanmar Name Entity Recognition and Transliteration
  - Myanmar Text Summarization
  - Myanmar Text Categorization

Overview of the System

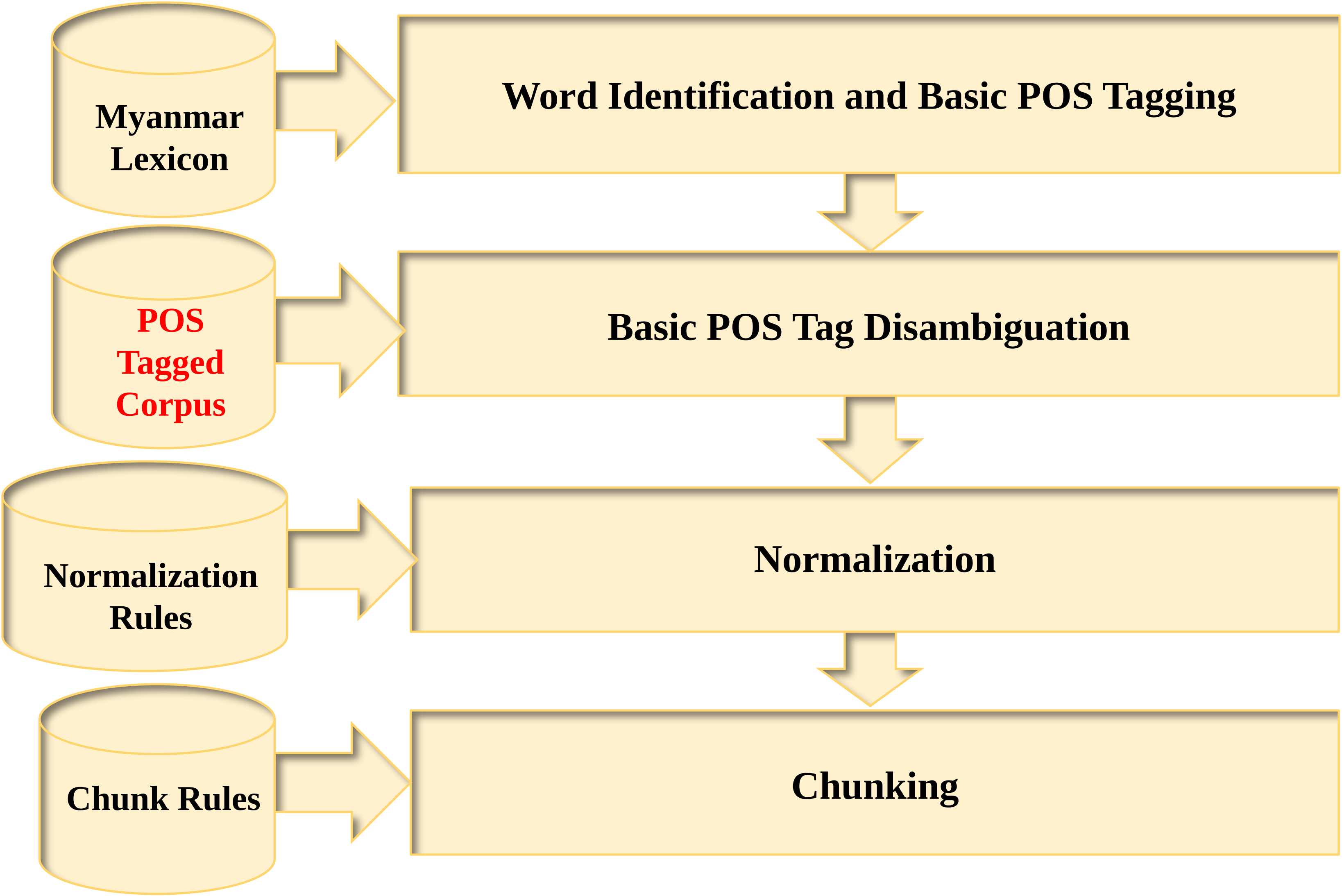


Source Language Analysis

- For Myanmar-English translation phase, it is the process of Myanmar Language Analyzer:
  - **Myanmar Part-of-Speech (POS) Tagging** and Chunking of Myanmar Language
  - Syntactic Analysis
    - Function Tagging and making Grammatical relation

- For English-Myanmar translation phase,
  - English POS and Chunking
  - Syntactic Analysis
    - Function Tagging and making Grammatical relation

Myanmar POS Tagging and Chunking



Myanmar Lexicon

Word Identification and Basic POS Tagging

POS Tagged Corpus

Basic POS Tag Disambiguation

Normalization Rules

Normalization

Chunk Rules

Chunking

Pre-tagged Corpus Format :

❑ Training Corpus

○ Myanmar words are segmented and tagged with their respective basic POS tags and categories as follows ::

✓ သူ/PRN.Person # ကျောင်း/NN.Building # သို့/PPM.Direction # သွား/VB.Common # သည်/SF.Declarative

✓ ကျောင်းသား/NN.Person # များ/Part.Number # ထဲတွင်/PPM.Extract # သူ/PRN.Person # အ/Part.Common # တော်/JJ.Dem # ဆုံး/Part.Common # ဖြစ်/VB.Common # သည်/SF.Declarative

✓ ဤ/PRN.Distobj # စာ/NN.Common # ကို/PPM.Obj # မည်သူ/PRN.Question # ရေး/VB.Common # ခဲ့ /Part.Support # သနည်း/SF.Interrogative

Example : Tagging

❑ Input Text

✓ သံလွင် မြစ် သည် မြန်မာပြည် တောင်ပိုင်း သို့ ဦးတည် စီးဆင်း သွား သည်။ (The river, Than Lwin, flows to south of Myanmar.)

❑ Tagging with All Possible Tags on Each Word

✓ သံလွင်_#NNP.Location

✓ မြစ် _#NN.Location

✓ သည် _#SF.Declarative #PPM.Subj

✓ မြန်မာပြည် _#NNP.Location

✓ တောင်ပိုင်း_#NN.Location

✓ သို့ _#PPM.Direction

✓ ဦးတည်_#VB.Common

✓ စီးဆင်း _#VB.Common

✓ သွား_#VB.Common#NN.Body#Part.Support

Disambiguation of Tags

- disambiguating all possible basic POS tags to produce the correct tag.

- training Myanmar pre-tagged Corpus with HMMs and LHMMs models.

- decoding using the Viterbi tagging algorithm to find out the best probable path (best tag sequence) for a given word sequence.

Example : Disambiguation

❑ **Disambiguation and Assigning with Correct Tag on Each Word**

✓	သံလွင်_#NNP.Location	(Than Lwin)
✓	မြစ် _#NN.Location	(The river)
✓	သည် _#PPM.Subj	(null)
✓	မြန်မာပြည် _#NNP.Location	(Myanmar)
✓	တောင်ပိုင်း_#NN.Location	(south)
✓	သို့ _#PPM.Direction	(to)
✓	ဦးတည်_#VB.Common	(flows)
✓	စီးဆင်း _#VB.Common	(flows)
✓	သွား_#Part.Support	(flows)
✓	သည် _#SF.Declarative	(null)

Example : Normalization

- forming more meaningful words and annotating with appropriate POS tags and categories.

❑ Before normalization,

"**ကျန်းမာ/VB.Common** # **ခြင်း/Part.Common** # သည် /PPM.Subj # လာဘ်/NN.Common # တစ်/NN.Cardinal # ပါး/Part.Type # ဖြစ်/VB.Common # သည် /SF.Declarative"

❑ After normalization,

"**ကျန်းမာခြင်း/NN.VBConvert** # သည် / PPM.Subj # လာဘ် / NN.Common # တစ် / NN.Cardinal # ပါး / Part.Type # ဖြစ်/ VB.Common # သည် / SF.Declarative "

Example : Chunking

- assemble the POS tagged words and identify chunk tag.

❑ Before chunking,

သူတို့/NNR.Person # သည်/PPM.Subj # အတန်း/NN.Common # ထဲတွင်/PPM.Extract # အတော်ဆုံး/JJS.Common # ကျောင်းသားများ/NNR.Person# ဖြစ်/VB.Common # ကြ/Part.Support # သည်/SF.Declarative

❑ After chunking,

NC [သူတို့/NNR.Person] # PPC [သည်/PPM.Subj] # NC [အတန်း/NN.Common] # PPC [ထဲတွင်/PPM.Extract] # **NC [အတော်ဆုံး/JJS.Common # ကျောင်းသားများ/NNR.Person] # VC [ဖြစ်/VB.Common # ကြ/Part.Support] # SFC [သည်/SF.Declarative]**

Alignment

- Identifying word correspondence that are translations of each other based on information found on parallel text.

- Developing a Myanmar-English bilingual corpus:
  - Dictionary lookup approach
  - Corpus-based approach

Word Alignment Algorithm

Step 1: Accept pair of Myanmar and English sentences.

Step 2: Tag English sentence with Part-Of-speech (POS) Tagger and it will produce tagged output also with root word.

Step 3: Segment Myanmar sentence into words.

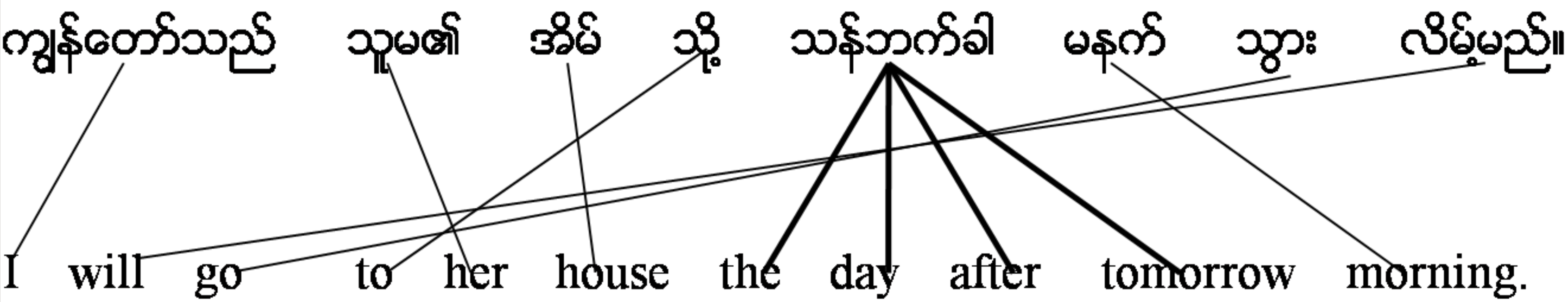
Removes the stop words.

Make morphological analysis of the noun and verb affixes using trigram method.

Step 4: Align the output English and Myanmar words from step 2 and 3 based on the first three IBM models and EM algorithm using parallel corpus.

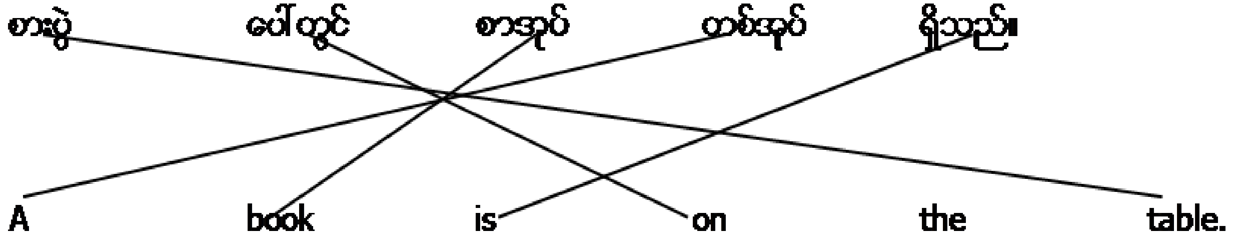
Step 5: Align the remaining words (i.e unaligned) using Myanmar-English bilingual dictionary.

Example Alignment



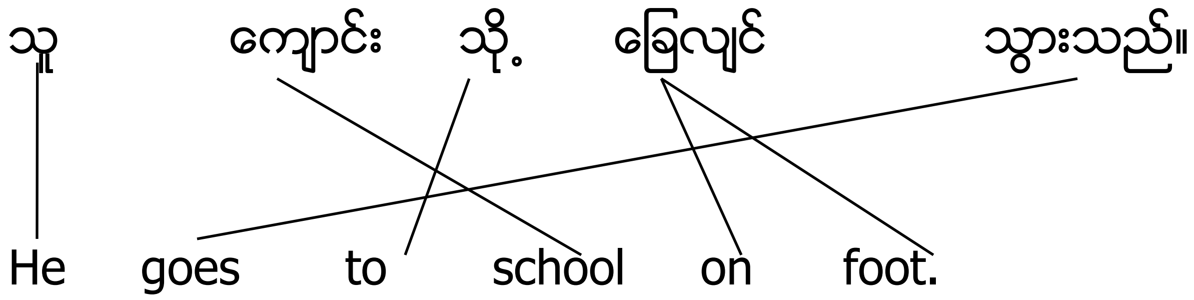
ကျွန်တော်သည် သူမ၏ အိမ် သို့ သန်ဘက်ခါ မနက် သွား လိမ့်မည်။

I will go to her house the day after tomorrow morning.



စားပွဲ ပေါ်တွင် စာအုပ် တစ်အုပ် ရှိသည်။

A book is on the table.



သူ ကျောင်း သို့ ခြေလျင် သွားသည်။

He goes to school on foot.

Problems in Alignment

- **Scarce Resource**
  - No publicly available POS-tagged corpus for Myanmar and English.
  - The constructed POS-tagged corpus has a limited number in size.

- **Linguistic Problem**
  - Parallel sentence pairs might not be equal size.
  - Myanmar and English word order could be significantly different.
  - Myanmar language is a morphologically rich and verb final language. English is a verb-second language.

Translation

- Phrase/word Translation pairs Extraction
- Morphological Analysis
- Word Sense Disambiguation

Phrase/word Extraction

- For each phrase we identified by its start position, end positions phrase length and target phrase to ensure that there are no gaps and no overlap.
- Applying N-gram methods using **Corpus**,

Source phrase	Start position	End position	Phrase Length	Target phrase	Translation probability
ငှက်	**1**	**1**	**1**	**Bird**	**1.0**
ငှက်များ	**1**	**2**	**2**	**Birds**	**1.0**
ပျံ	**4**	**4**	**1**	**Fly**	**1.0**
ပျံကြသည်	**4**	**6**	**3**	**Fly**	**1.0**

- **Translation**

ငှက်များ	-	birds
ပျံကြသည်	-	fly

Example : Morphological Analysis of verbs

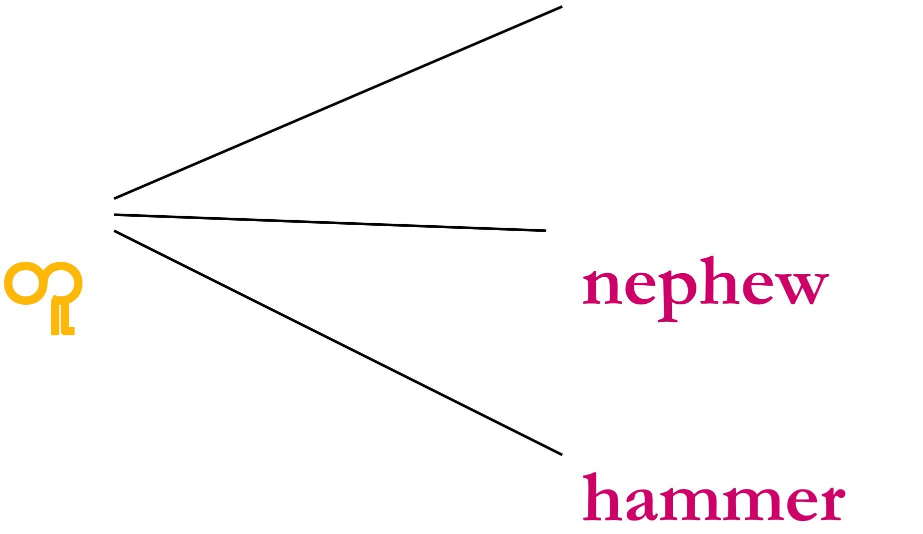
- Myanmar unknown verb: ကြည့်ခဲ့ပါသည်
- Main Verb: ကြည့်
- Verb suffiex: ခဲ့ပါသည်
- Tense particle: ခဲ့
- Translation of main verb (using corpus): look
- Generation of surface word: ကြည့်/look, ခဲ့/past ပါသည်/null(suffix)
- ကြည့်ခဲ့ပါသည်/looked

Word Sense Disambiguation for Myanmar Language

- Purpose:
  - to solve the ambiguity of Myanmar words for Myanmar-English machine translation

Ambiguous Example

- Noun Examples



တူ

chopsticks

nephew

hammer

- သူသည်တူဖြင့်ခေါက်ဆွဲစားသည်။ He eats the noodle with chopsticks.
- သူ့မှာတူသုံးယောက်ရှိသည်။ He has three nephews.
- လက်သမားသည်တူကိုသုံးသည်။ Carpenter uses the hammer.

WSD Algorithm for Myanmar Word

Step1:Preprocessing

-Segment input sentence

-Remove stop words from input sentence and create ambiguous vector

Step2:Multi-sense Look-up

-Retrieve all possible sense meanings of ambiguous word from corpus

-Collect training data concerning with these sense from corpus

Step3:Build context vectors for each sense based on collected training data

-For all context vectors do

  -Remove stop words

  -Remove redundant words

-End For

Step4:Calculate the cosines between ambiguous vector and each of the context vectors

$$\cos\theta = \frac{A.B}{\|A\|.\|B\|} = \frac{\sum_{i=1}^{n} A_i.B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}.\sqrt{\sum_{i=1}^{n} B_i^2}}$$

where A represents each word in ambiguous vector

B represents each word in each context vector

Step5:Choose correct sense of the target word

$s' = \text{argmax score}(s_i)$

Conclusion

- The data sparseness is most important in many research regarding NLP because of the followings:
  - The rules only can not be solved for all problems for many languages.
  - So, the researches are coming based on the statistical model.
  - The more availability of data in developing the system/tools, the more accuracy we can get.
- So, ALT data is very useful not only for Myanmar language but also for all languages to be applied in various kinds of NLP researches.

Thank you!