10th International Conference on Science and Engineering 2019

Yangon Technological University (YTU), Myanmar

Rule-based Myanmar Language Chatbot for Travel and Tourism Domain

8.12.2019 (Sunday)

Htet Htet Yadanar Oo[1], May Thae Sue[2], Ye Kyaw Thu[3,4], Hnin Aye Thant[1], Nang Aye Aye Htwe[2]

[1]University of Technology (Yatanarpon Cyber City), Myanmar
[2]Mandalay Technological University, Myanmar
[3]National Electronics and Computer Technology Center (NECTEC), Thailand
[4]Language Understanding Lab., Myanmar

Outlines

- Introduction

- Aim and Objectives

- Data Preparation and Collection

- Regular Expression (RE)

- Word Segmentation

- POS Tagging

- String Similarity

- System Design

- System Implementation

- Evaluation

- Features and Limitations

- Future Extensions

Introduction

- Chatbot is a computer program that can communicate with human through natural language via text or voice input.

- Artitificial Intelligence (AI) or Natural Language Processing (NLP) is usually embedded in a chatbot to understand the context and perform much more complicated tasks than what a bot does

- Travel and tourism chatbot is a human-computer dialog system that operates through natural language via text.

- This proposed system is a virtual assistant programmed to automatically answer user requests and can perform as a travel adviser for local travelers

- It can provide 24-7 customer services and support.

Aim and Objectives

Aim

- To help the users who want to travel around Myanmar

Objectives

- To provide transportation advices for the users

- To give suggestion and recommendation places for travellers

- To suggest suitable moments for the trip

- To make local people easy and comfortable for their trip

Data Preparation and Collection

- Prepare and collect parallel Myanmar sentences (Questions and Answers).

- Collect **7394** parallel Myanmar sentences for Mon State, Kachin State, Shan State, Mandalay and Pyin Oo Lwin regions

For example:

Q - မွန် ပြည်နယ် မှာ အထင်ကရ နေရာ တွေ က ဘယ် နေရာ တွေ လဲ ။

A - မွန် ပြည်နယ် မှာ အထင်ကရ နေရာ တွေ က ကျိုက်ထီးရိုး ဘုရား ၊ ကျိုက္ခမီ ရေလယ် ဘုရား တို့ ဖြစ်

တယ် ။

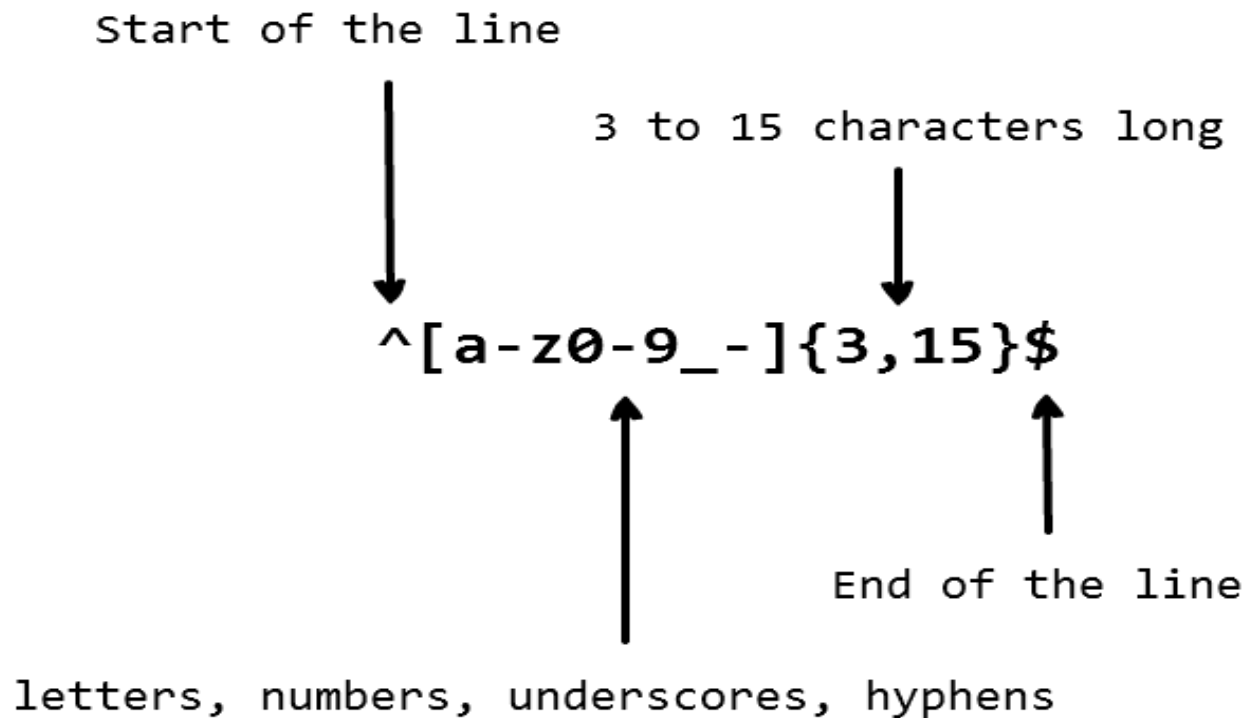Q - မန္တလေး မှာ ဘာ ဘုရား တွေ ရှိ လဲ ။

A - မဟာမြတ်မုနိ ဘုရား ၊ မန္တလေး တောင် တို့ ရှိ ပါတယ် ။

Q - ကချင် ပြည်နယ် ရဲ့ မြို့တော် က ဘယ် မြို့ လဲ ။

A - ကချင် ပြည်နယ် ရဲ့ မြို့တော် က မြစ်ကြီးနား မြို့ ဖြစ် ပါ တယ် ။

Regular Expression (RE)

- **RE** is a special sequence of characters that help **match** or **find** other strings using specialized pattern

- Build **RE** pattern with python **re module**

```
Start of the line

                        3 to 15 characters long

        ^[a-z0-9_-]{3,15}$

                                End of the line

letters, numbers, underscores, hyphens
```

Cont'd

- Parallel corpus that we created in the database

  Q - **မွန်ပြည်နယ်** မှာ �‌ဘာ **ဘုရား** တွေ ရှိ လဲ။

  A - မွန်ပြည်နယ် မှာ ကျိုက်ထီးရိုး ဆံတော်ရှင် ‌စေတီ၊ ကျိုက္ခမီ ဘုရား ... တို့ ရှိ ပါတယ် ။

  Q - **မွန်ပြည်နယ်** မှာ ‌ဘာ **ကမ်းခြေ** တွေ ရှိ လဲ။

  A - မွန်ပြည်နယ် မှာ စက်စဲကမ်းခြေ၊ ကဗျာဝကမ်းခြေ ... တို့ ရှိ ပါတယ် ။

- Regular Expression pattern

  - r'(.***မွန်ပြည်နယ်**.***ဘုရား**.*)'
  - r'(.***မွန်ပြည်နယ်**.***ကမ်းခြေ**.*)'

  User Input - **မွန်ပြည်နယ်** မှာ ရှိတဲ့ **ဘုရား** တွေ ကို သိ ချင် ပါတယ် ။

  Chatbot - မွန်ပြည်နယ် မှာ ကျိုက်ထီးရိုး ဆံတော်ရှင် ‌စေတီ၊ ကျိုက္ခမီ ဘုရား ... တို့ ရှိ ပါတယ် ။

  User input - **မွန်ပြည်နယ်** မှာ ရှိတဲ့ **ကမ်းခြေ** တွေ ကို အလည် သွား ချင် ပါတယ် ။

  Chatbot - မွန်ပြည်နယ် မှာ စက်စဲကမ်းခြေ၊ ကဗျာဝကမ်းခြေ ... တို့ ရှိ ပါတယ် ။

Word Segmentation

- In Myanmar text, spaces are not usually inserted for word boundaries

- Words are composed of single or multiple syllables, not usually separated by white spaces

- **Conditional Random Fields (CRFs)** is used to develop segmentation model

- To segment Myanmar Sentences using CRF Suite
  - ➔ Break sentence into **syllable word**
  - ➔ Tag each syllable words with **segmentation tags** (labeling)
  - ➔ Train the segmentation model using **Conditional Random Fields (CRF suite)**

Cont'd

Tag number	Tag	Position
1	<	The first syllable/character in a word
2	-	Represents both < and |
3	|	Final syllable/character in a word

Unsegmented input: မွန်ပြည်နယ်ရဲ့မြို့တော်ကမော်လမြိုင်မြို့ဖြစ်ပါတယ်။

Syllable breaked sentence: မွန်_ပြည်_နယ်_ရဲ့_မြို့_တော်_က_မော်_လ_မြိုင်_မြို့_ဖြစ်_ပါ_တယ်_။

Tag set for CRF Suite: \| <   \| \| <   \| \| <   -   \| \| \| \| \| \|

Segmented output: မွန်_ပြည်နယ်_ရဲ့_မြို့တော်_က_မော်လမြိုင်_မြို့_ဖြစ်_ပါ_တယ်_။

9

POS Tagging (Part of Speech)

- POS tagging is used to extract keywords from the user input and to measure string similarity between user input and existing question in database

- Only keywords (noun, verb, adj, adv) are selected and all other words (part, ppm, conj and etc) are dropped

- The input sentence is tagged with the following POS tag

- Abbreviation(abb) → အထက (Basic Education High School)

- Adjective(adj)→ ရဲရင့် (brave), လှပ (beautiful)

- Adverb(adv) → ဖြေးဖြေး (slow), နည်းနည်း (less)

- Conjunction(conj) → နှင့် (and), သို့မဟုတ် (or)

- Foreign Word(fw) → 1, 2, 3, Myanmar, BBQ

- Verb(v) → ကူညီ (help), သွား (go), စား (eat)

- Noun(n) → ကျောင်း (school), စာအုပ် (book)

- Number(num) → ၁ (1), ၂ (2), ၃ (3)

- Particle(part) → များ (plural nouns as "-s", "-es"), လိမ့် (will)

- Post-positional Marker(ppm) → သည်, က, ကို, သို့, (at,on,in,to)

- Pronoun(pron) → ကျွန်တော် (I), ကျွန်မ (I), သင် (You), သူ (He)

Cont'd

- Unsegmented sentence:

  မွန်ပြည်နယ်ရဲ့အထင်ကရဘုရားတွေကိုသိချင်ပါတယ်။

- Segmented sentence:

  မွန်_ပြည်နယ်_ရဲ့_အထင်ကရ_ဘုရား_တွေ_ကို_သိ_ချင်_ပါ_တယ်_။

- POS tagged sentence:

  မွန်/n_ပြည်နယ်/n_ရဲ့/ppm_အထင်ကရ/adj_ဘုရား/n_တွေ/part_ကို/ppm_သိ/v_ချင်/part_ပါ/part_တယ်/ppm

- Keyword for string similarity: မွန်_ပြည်နယ်_အထင်ကရ_ဘုရား_သိ

String Similarity

- Compare a set of two strings that how much these two strings be closed

- Use **Damerau-Levenshtein Distance algorithm**

- Find minimum distance between the two strings are with these operations; **deletion, insertion, substitution and transposition**

  - Insertion   eg. cat > cart

  - Deletion   eg. cart > cat

  - Substitution   eg. cart > dart

  - Transposition   eg. water > watre

- Distance ('a cat','an cat') → 1

- Distance ('a cat','an act') → 2

Cont'd

Parallel corpus

Q - ရေးမြို့ မှာ ဘာ ကမ်းခြေ တွေ ရှိ လဲ။

A - ရေးမြို့ မှာ ကဗျာဝကမ်းခြေ၊ ပင်လယ်ဝကမ်းခြေ တို့ ရှိ ပါတယ်။

Q - ရေးမြို့ မှာ ဘာ ဘုရား တွေ ရှိ လဲ။

A - ရေး မြို့ မှာ လေးမျက်နှာ ဘုရား ကြီး၊ ကျိုက်မင်းပလော့ ဆံတော်ရှင် စေတီ ... တို့ ရှိ ပါတယ်။

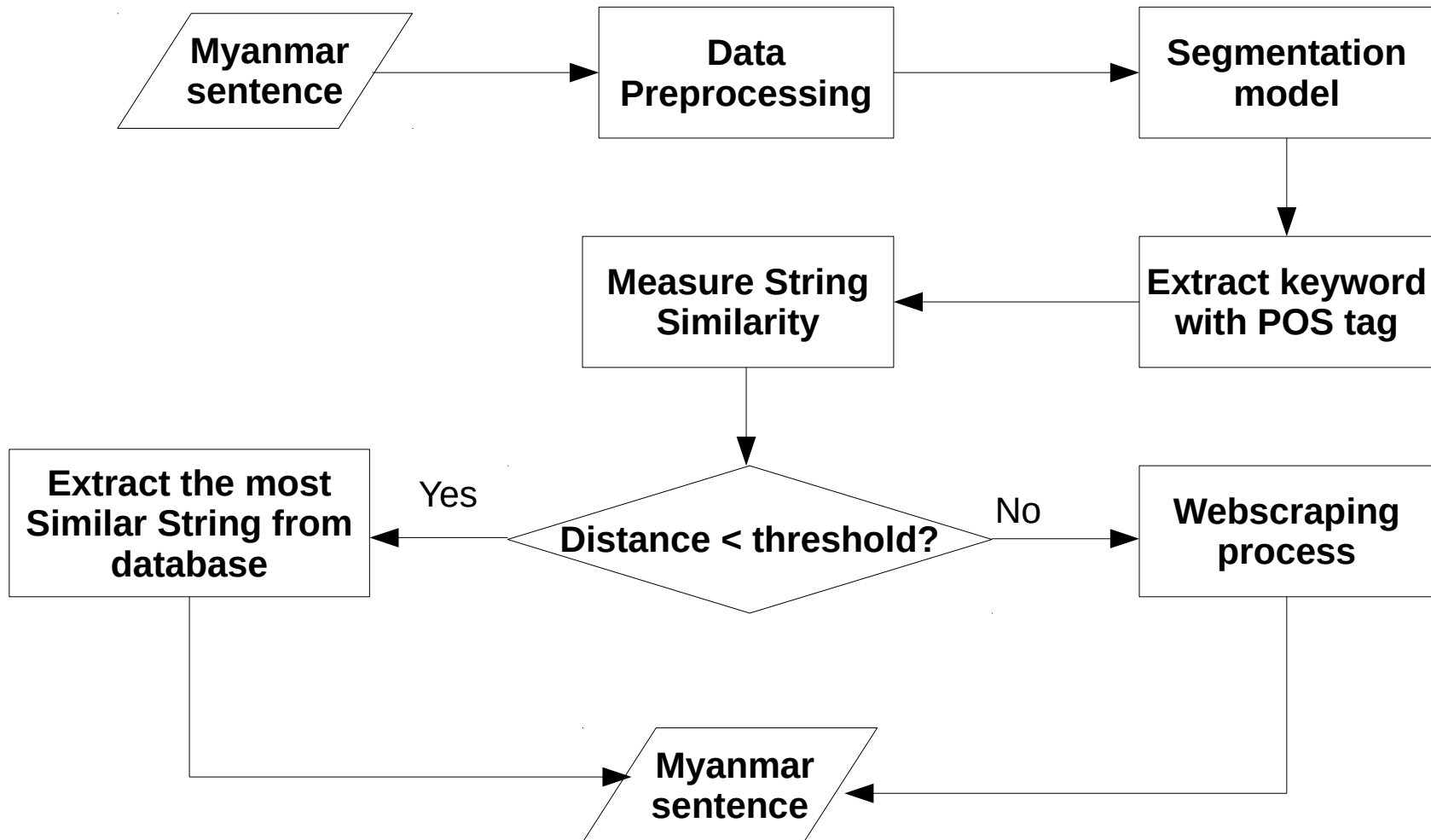**Example of how similar two strings and the extracted keywords from the user input and existing question in database**

User input – ရေးမြို့ မှာ ကမ်းခြေ တွေ ကို လည် ချင် တယ် ။

Distance ("ရေးမြို့_မှာ_ဘာ_ကမ်းခြေ_တွေ_ရှိ_လဲ။" , "ရေးမြို့ _ကမ်းခြေ_လည်″) → 17

Distance ("ရေးမြို့_မှာ_ဘာ_ဘုရား_တွေ_ရှိ_လဲ။" , "ရေးမြို့ _ကမ်းခြေ_လည်″) → 23

Chatbot Reply - "ရေးမြို့ မှာ ကဗျာဝကမ်းခြေ၊ ပင်လယ်ဝကမ်းခြေ တို့ ရှိ ပါတယ်။″

System Design

System Implementation

- Python 3.7.3

- Flask framework

Python Library

- Python **re** module

- Beautifulsoup4

- Urllib2

- request

- google_images_download

Google Map Implementation

- Google map feature is used to search and reply google map when the users ask questions about the **location or direction**

- Regular expression pattern to check if the user asks for the location or direction:

  - .*လမ်းကြောင်း.* or .*လမ်းညွှန်.* or .*�‌�‌ဘယ်.*မှာ.*

- User input – ပြင်ဦးလွင် မြို့ ကို **လမ်းညွှန်** ပေး ပါ ဦး ။
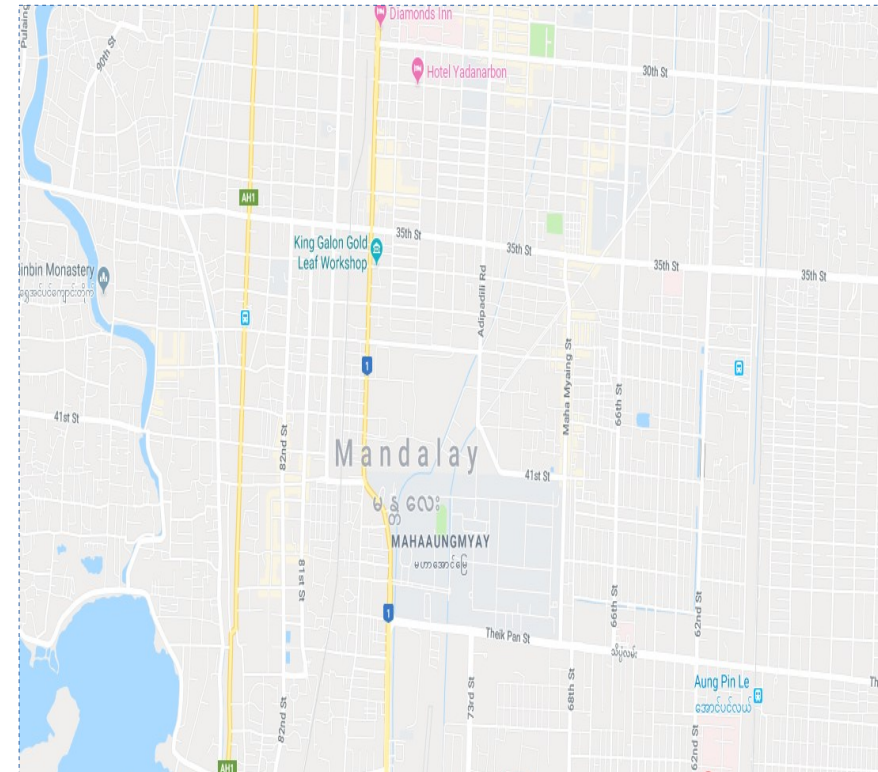
YouTube Video Implementation

- Chatbot can reply both text message and videos from YouTube that the user can be satisfied.

- Beautifulsoup4 and urllib libraries are used to implement YouTube Videos

- Example of search query -

    Url → "https://**www.youtube.com**/results?search_query=" + **query**

- User input – ကချင် ပြည်နယ် က လု လား ။

- Chatbot – Scrape **YouTube video** that is corresponding with the user search input

Web Scraping Link, Image for out of Database State

- Chatbot will also response search results with corresponding website link , description and images about the user input from Google
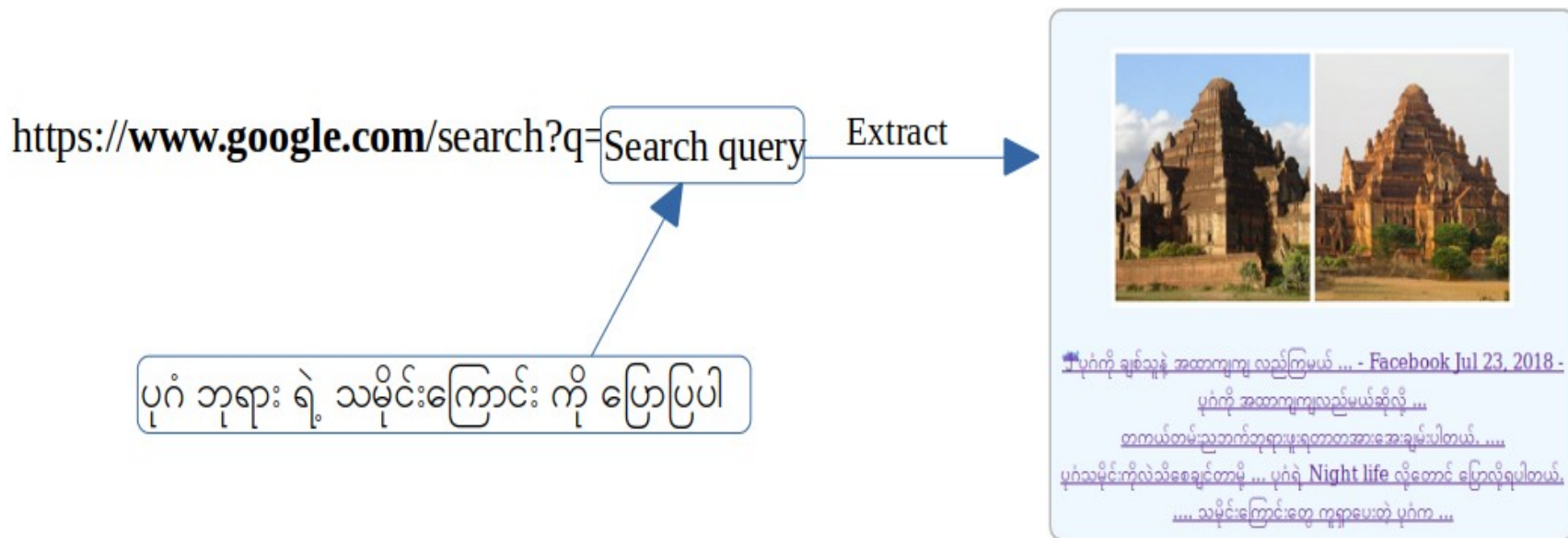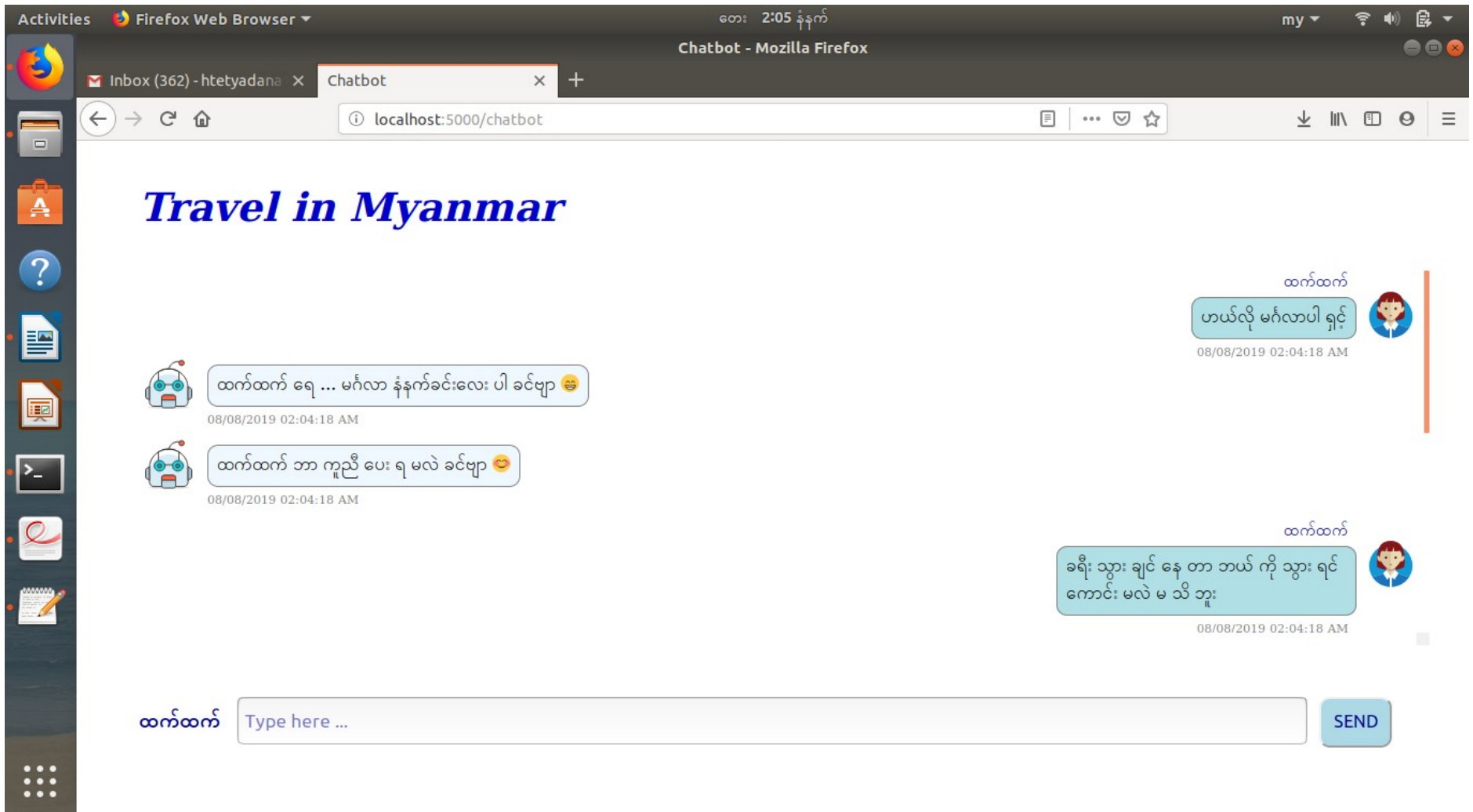
Scraping Hotel Link

- Chatbot can reply hotel link, description and price of hotel

- Regular Expression pattern to check if the input is to search Hotel

  - .*ဟိုတယ်.*

  - .*ဟော်တယ်.*

  - .*တည်းခိုခန်း.*

  - .*hotel.

  - .*မိုတယ်.*

- User input – **နေပြည်တော်** မှာ �‌ဘာ **ဟိုတယ်** တွေ ရှိ လဲ။

- Url → https://www.google.com/search q=**နေပြည်တော်+hotel**
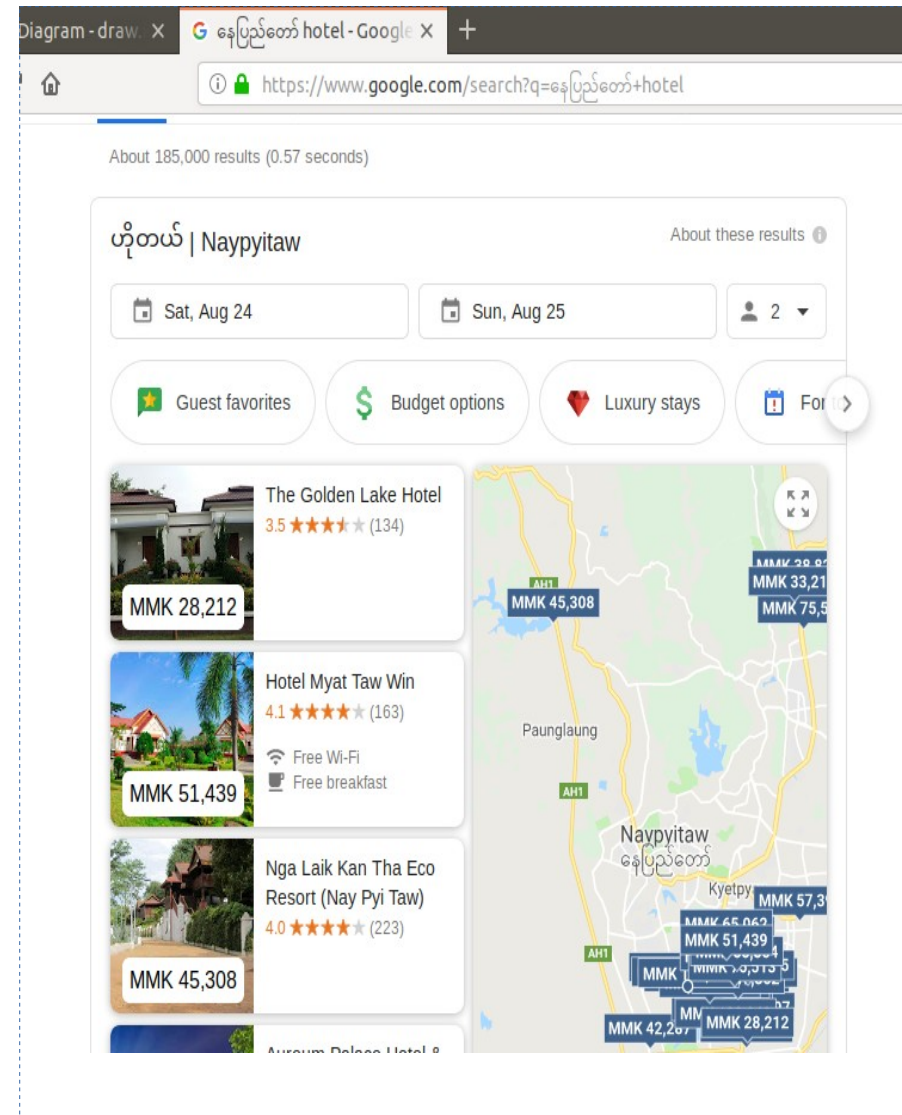
System output

Evaluation

- A Likert scale is a rating scale that measures how people feel about products, services, or experiences

- Eight user satisfaction survey questions are prepared and each question was defined with satisfaction range 1 to 5

- Ten conversation scenarios are created based on several categories such as pagodas, hotels, beaches, etc

- This system is evaluated with 16 users (13 females and 3 males), the age range is between 22 and 26

Overall Likert scale evaluation

	User 1	User 2	User 3	User 4	User 5	User 6	User 7	User 8	Overall
Easy to use	5	3	3	4	3	3	4	5	3.75
Easy to understand	3	3	3	2	3	3	3	3	2.88
Fast to reply	3	4	5	4	3	4	3	4	3.75
Satisfied on response	2	3	3	3	2	2	3	3	2.63
Expected response	3	3	3	3	2	2	2	3	2.63
Give help and advice	3	3	2	3	3	3	3	3	2.88
Resonable response	3	2	3	2	3	2	2	3	2.5
Future use	3	3	3	4	4	3	3	4	3.38

Features and Limitations

FEATURES

- Give informations about pagoda, popular events, famous places, restaurants, markets, hotels, beaches, waterfalls, monastery, weather, transportation

- Reply Google Links, images, YouTube videos, Google map direction when the user input is not in created database

LIMITATIONS

- Only questions about Mon State and Mandalay region can be replied because only data of these two places have been cleaned and structured

- Only Myanmar Language is supported

Future Extensions

- To prepare and collect more detailed informations about places in Myanmar

- To develop voice recognition between user-chatbot interaction

- To be enable to translate conversation of users and chatbot to English Language to support for the foreigners

- To develop chatbot interface on Messenger platform

Thank You …

QUESTIONS?