# Myanmar Named Entity Recognition with a Statistical Approach

# Named Entity Recognition

- Named-Entity Recognition (NER) is a subtask of information extraction that seeks to locate and classify named entities in text into predefined categories such as person names, time expression, organisation, locations etc.

- Myanmar Natural Language Processing (NLP) field is at developing stage and data resources are still limited.

- A statistical HMM(Hidden Markov Model) based model is used to implement our system.

- The output of the system is to identify NE tagged data.

# Previous NER Work

**Hsu Myat Mo and Khin Mar Soe, "Syllable-based Neural Named Entity Recognition for Myanmar Language”**

- In this paper, a baseline CRF-based statistical NER model for Myanmar language is proposed as the main objective of the system.

- This system recognises the six types of named entities.

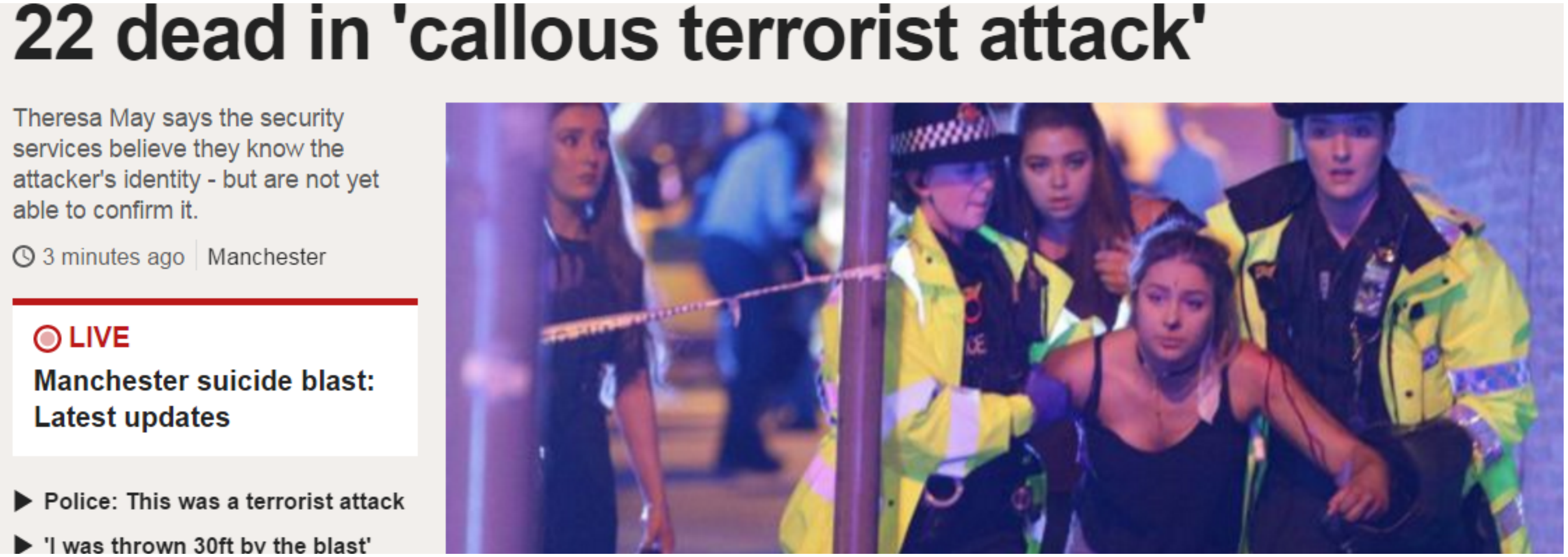
- This paper focused on to learn NER problem as sequence learning problem and to introduce a deep neural architecture for Myanmar NER Modelling.

# Example Applications of NER

- **Classifying content for news providers**
- Demo Text: ကြက်တောင်အားကစားမယ် သက်ထားသူဇာ မလေးရှားသို့ သွားရောက် လေ့ကျင့်မည်
- Keywords :
  - Person Name - သက်ထားသူဇာ
  - Location - မလေးရှား

# Example Applications (Cont’d)

- **Powering Content Recommendations**



22 dead in 'callous terrorist attack'

Theresa May says the security services believe they know the attacker's identity - but are not yet able to confirm it.

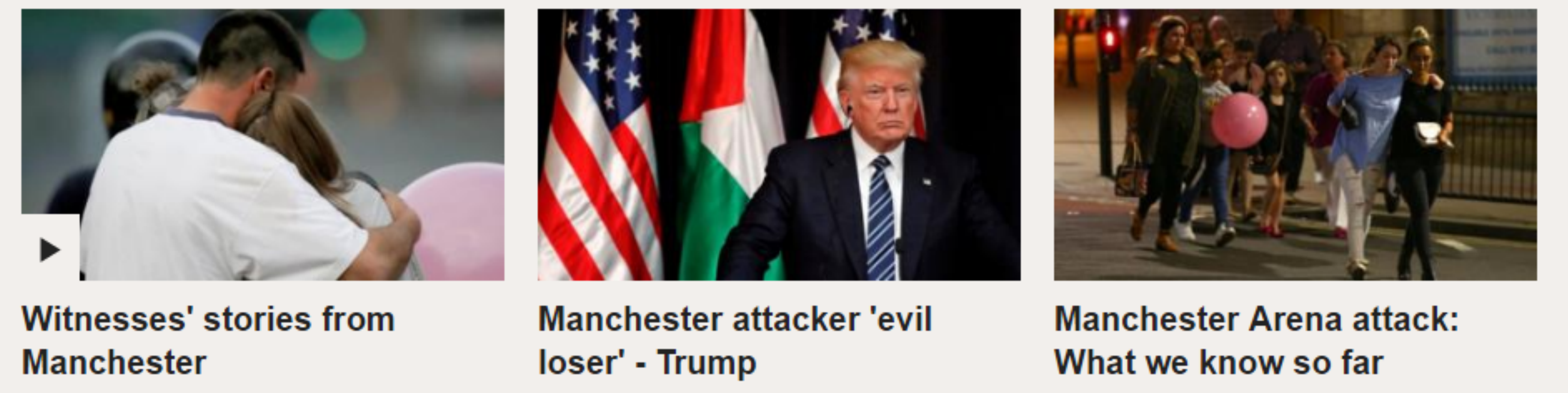
3 minutes ago | Manchester

LIVE

**Manchester suicide blast: Latest updates**

▶ Police: This was a terrorist attack

▶ 'I was thrown 30ft by the blast'

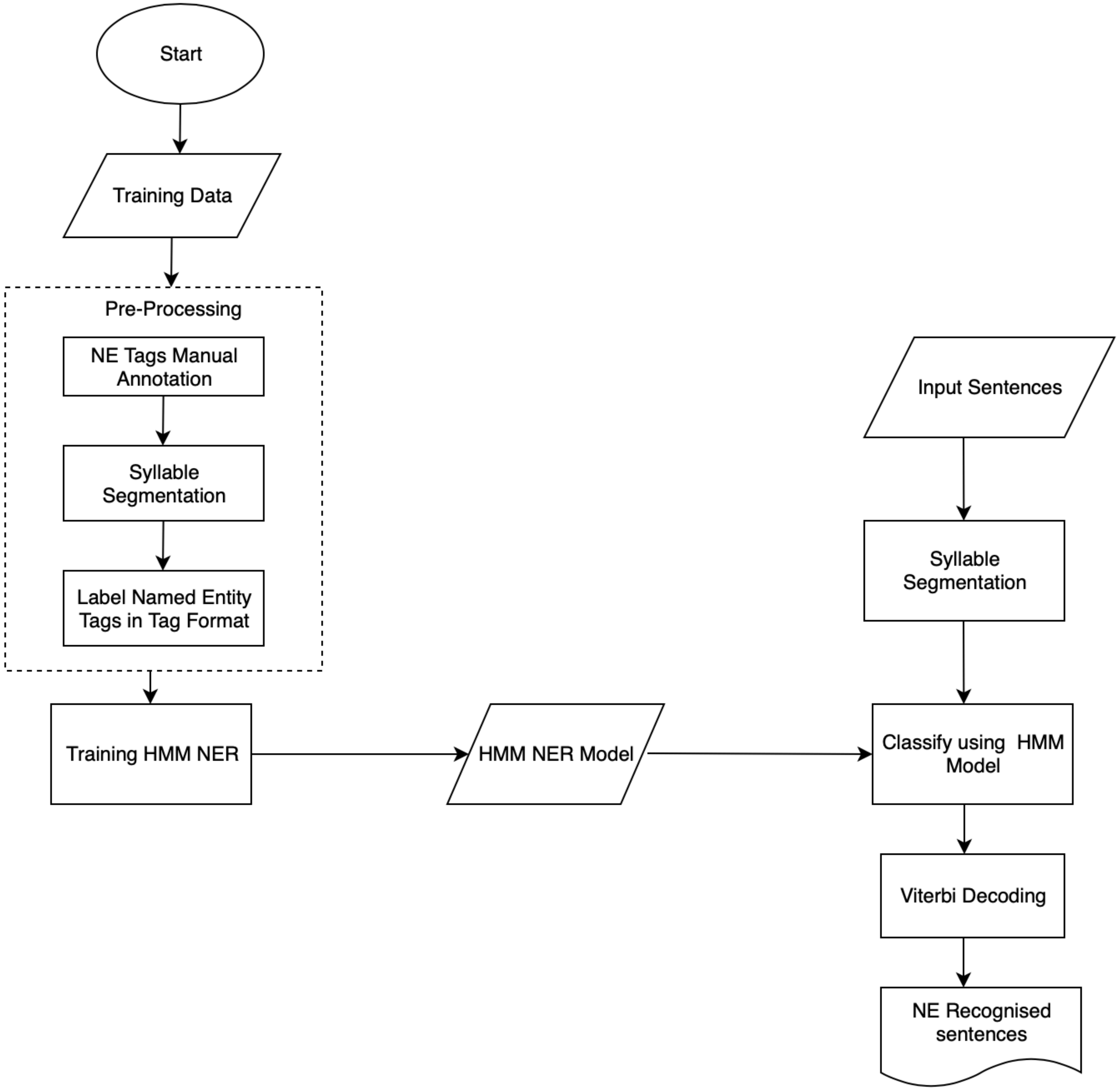


Witnesses' stories from Manchester

Manchester attacker 'evil loser' - Trump

Manchester Arena attack: What we know so far

# System Flow



**System Flow of the Proposed System**

# Dataset

- The data is developed from the leading Myanmar Newspaper available in the website. We can get huge amount of data from news website.

- In Myanmar NER corpus, news sentence written with Myanmar scripts ranging from 2016 to 2020, from online official news website such as 7days Daily news, Eleven media news, Myanmar Now and so on are collected and used.

- Different types of news gender such as business, crime, health, education, technology, sport, religion, environment, and also politics are included.

# Data Preparation

- As a first step, data cleaning is carried out to develop the system.
- Mistyped errors are included in the news articles.
- Wrongly typed error of ရ and ၇ such as ရန်ကုန် / ၇န်ကုန် is found frequently in the news.
- Therefore, all types of errors are corrected manually.
- It is essential to correct wrongly typed error because the quality of data strongly affects the performance.
- All collected data are standard Unicode encoding.

# Defining NE Tags

- Each NE has to be annotated with NE tag to indicate Name Entities in sentences.
- In this system, there are five major types of NE tags that are pre-defined for manual annotation: **PNAME, LOC, ORG, TIME, NUM**
    - **PNAME** - to indicate person names
    - **LOC** - to indicate location entities
    - **ORG** - to indicate names of organisations
    - **TIME** - to indicate dates, months, and years
    - **NUM** - to indicate number format
    - **OTHER** - to indicate others

# Defined NE Types and their Usage

| Defined NE Types | Example Usage |
|---|---|
| PNAME | အောင်ဆန်းစုကြည်၊ အေးမြဖြူ၊ ဒေါ်နယ်ထရန့် |
| LOC | ရန်ကုန်၊ နယ်သာလန်၊ ရွှေတိဂုံ |
| ORG | ရန်ကုန်ကွန်ပျူတာတက္ကသိုလ်၊ ပညာရေးဝန်ကြီးဌာန |
| TIME | ဇန်နဝါရီ၊ စနေ၊ ၁၇.၈.၁၉၉၆ |
| NUM | ၁၇၊ ၁၀၀၀၀၊ ၂.၁၇ |
| OTHER | အခြားအရာများ |

# Example sentences from Myanmar NE Tagged Corpus

- **ဒီဇင်ဘာ** @TIME| **၈** @TIME| ရက် **ရန်ကုန်**@LOC| တွင်ဒေါ်@OTHER| **အောင်ဆန်းစုကြည်** @PNAME| ထောက်ခံပွဲကိုလူအင်အား@OTHER| **၂၅၀၀၀**@NUM| ကျော်ပါဝင်ခဲ့သည်။ @OTHER|

- **လှိုင်သာယာ**@LOC| တွင်ယာဉ်@OTHER| **၉**@NUM| စီးဆင့်တိုက်ပြီးလူ@OTHER| **၈**@NUM| ဦးဒဏ်ရာရခဲ့သည်။@OTHER|

- **စင်ကာပူ**@LOC| ၊@OTHER| **တရုတ်**@LOC| ၊@OTHER| **ဟောင်ကောင်**@LOC| ၊@OTHER| **ဗီယက်နမ်**@LOC| နှင့်@OTHER|**ဂျာမနီ**@LOC| တို့မှ@OTHER| **ရန်ကုန်**@LOC| တိုင်း အတွင်းဒေါ်လာ@OTHER| **၂၉**@NUM| သန်းကျော် ရင်းနှီးမြှုပ်နှံမည်။@OTHER|

# Tagging Scheme

- As tagging scheme, IOBES scheme is supposed for experiments.
    - I - An inner token of a multi-token entity
    - O - A non-entity token
    - B - the first token of a multi-token entity
    - E - the final token of a multi-token entity
    - S - the single token

# Example of Tagging Scheme

- ဒီဇင်ဘာ
  - ဒီ  B-TIME
  - ဇင် I-TIME
  - ဘာ E-TIME

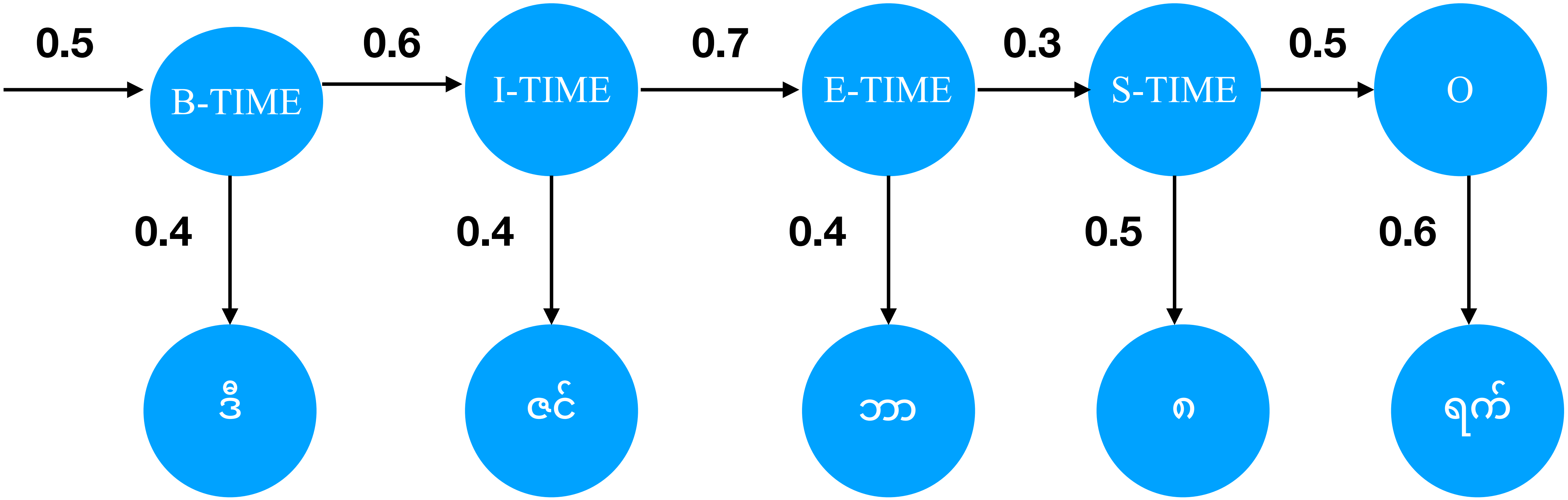
- သီရိရွှေစင်
  - သီ B-PNAME
  - ရိ I-PNAME
  - ရွှေ I-PNAME
  - စင် E-PNAME

- ဝင်ရောက်သည်
  - ဝင် - O
  - ရောက် - O
  - သည် - O

# Hidden Markov Model (HMM)

- HMM is a successful model in various sequence labelling tasks.
- It is a type of generative model and based on Markov chain and also, each label is represented as states.
- HMM model defines the joint probability for each observation symbol and the state transition.
- In HMM training, trigram model will be proposed to develop the system.

# Hidden Markov Model (HMM) (Cont’d)



# Decoding

- The test of decoder is to find the best hidden state sequence given an HMM input and a sequence of observations.

- Viterbi decoding is the most common decoding algorithm used for HMM based tagging task.

- It makes HMM more efficient on the decoding of Named Entities.

- In this stage, the Viterbi algorithms use to search the best word sequence.

# Example of Calculating Performance

- **Manual Annotation** - **ဂမ်ဘီယာ**နိုင်ငံ ICJ တွင် သွားရောက် တရားရင်ဆိုင်မည့် နိုင်ငံတော် အတိုင်ပင်ခံ ဒေါ် **အောင်ဆန်းစုကြည်** အားပေးထောက်ခံသည့်အနေဖြင့် **မန္တလေး** တွင် လူအင်အားများစွာဖြင့် လူထုထောက်ခံပွဲ ကျင်းပ

- **Actual Extraction from System** - **ဂမ်ဘီယာ**နိုင်ငံ ICJ တွင် သွားရောက် တရားရင်ဆိုင်မည့် နိုင်ငံတော် အတိုင်ပင်ခံ ဒေါ် **အောင်ဆန်းစုကြည် အားပေး**ထောက်ခံသည့်အနေဖြင့် မန္တလေးတွင် လူ**အင်အား**များစွာဖြင့် လူထုထောက်ခံပွဲ ကျင်းပ

- True Positive - ဂမ်ဘီယာ၊ အောင်ဆန်းစုကြည်

- False Positive - အားပေး၊ အင်အား

- False Negative - မန္တလေး

- Precision - 2/2+2 = 2/4 = 0.5

- Recall - 2/2+1 = 2/3 = 0.67

- F-1 Score - 2 * ( (0.5*0.67) / (0.5 + 0.67)) = 0.5726

# Conclusion

- Named Entity Recognition, also known as entity extraction classifies named entities that are present in a text into pre-defined categories like “individuals”, “companies”, “places”, “organisation”, etc.

- It adds a wealth of semantic knowledge to our content and helps us to promptly understand the subject of any given text.

- NER systems have been created that use linguistic grammar-based techniques as well as statistical models such as machine learning.

# You can reach me

- Facebook (Tin Latt Nandar)
- Gmail (tinlattnandar.cs@gmail.com)
- LinkedIn (Tin Latt Nandar)

# Q & A Session

# Thank You