ICB Milestone-1 Report

– Group Member: R04922029 薛祐婷、R05922062楊宗翰、R05944004 巫承威、R05922136 黃博政、R05922131 祝子軒
– Demo site: http://ntu-course-chatbot.ml (http://ntu-course-chatbot.ml/)

- ## Ontology Construction

1. Backend Database

- **Backend Engine**
  - Python 3.5.2
  - Django 1.10.6
  - SQLite
- **Model**
  - Course (crawled from NTU Online台大課程網 (https://nol.ntu.edu.tw/nol/guest/index.php))
  - Review (crawled from PTT - NTUcourse board)

2. Dialogue Semantic Schema

- **Supported Tables/Domains**
  - 2 tables (similar with the table on NTU Online台大課程網 (https://nol.ntu.edu.tw/nol/guest/index.php))
    - Course
    - Review
  - 1 domain (only NTU's course related)
- **Supported User Intents**
  - Search for Course (including **4** intents)
    1. **classroom**: 自然語言處理在哪裡上課?
       request.query_course(goal='classroom', title='自然語言處理')
       -> '資105'
    2. **instructor**: 機器學習有哪些老師開課?
       request.query_course(goal='instructor', title__contains='機器學習') -> ['李宏毅', '林智仁', '林軒田']
    3. **schedule**: 機器學習技法是在什麼時候上課?
       request.query_course(goal='schedule', title='機器學習技法')
       -> '二5,6'
    4. **title**: 陳信希老師有開什麼課?
       request.query_course(goal='title', instructor='陳信希')

-> [‘專題研究’, ‘自然語言處理’]

- The support for review searching will be added later

- **Supported Slots**
  - when
  - title
  - instructor

3. Generated Training Examples for each Intent

- **Random Template Components**
  - **pihua_start**
    [‘’, ‘請問’, ‘告訴我’, ‘我想知道’, ‘幫我找’, ‘幫我查’, ‘幫我查一下’ , ‘查一下’]
  - **pihua_end**
    [‘’, ‘謝謝’, ‘感謝’, ‘感恩’, ‘謝謝你’]
  - **where**
    [‘在哪’, ‘哪裡’, ‘在哪裡’, ‘哪間教室’, ‘在哪間教室’, ‘在什麼地方’, ‘哪個館’, ‘哪個系館’, ‘在哪個系館’, ‘在哪個系館哪間教室’]
  - **question**
    [‘’, ‘嗎’, ‘你知道嗎’]
  - **when**
    [‘’,‘今天’,‘星期一’,‘星期二’,‘星期三’,‘星期四’,‘星期五’,‘禮拜一’,‘禮拜二’,‘禮拜三’,‘禮拜四’,‘禮拜五’]
  - **course_query**
    [‘有哪些課’, ‘開哪些課’, ‘有開哪些課’ ,‘教哪些課’]
  - **instructor_query**
    [‘有哪些’, ‘是哪個’, ‘是哪位’]
  - **time_query**
    [‘什麼時候’, ‘在幾點’, ‘在星期幾’, ‘在禮拜幾’, ‘幾點幾分’]
  - **teacher**
    [‘老師’, ‘教授’]

- **classroom**
  - {{pihua_start}}，{{when}}{{title}}{{where}}上課{{question}}，{{pihua_end}}?
  - {{pihua_start}}，{{when}}{{instructor}}的{{title}}{{where}}上課{{question}}，{{pihua_end}}?

- **instructor**
  - {{pihua_start}}，{{title}}有誰開課，{{pihua_end}}?
  - {{pihua_start}}，{{title}}有誰上課，{{pihua_end}}?

- {{pihua_start}}，{{title}}是誰開課，{{pihua_end}}?
    - {{pihua_start}}，{{title}}是誰上課，{{pihua_end}}?
    - {{pihua_start}}，{{title}}{{instructor_query}}{{teacher}}，{{pihua_end}}?
- **schedule**
    - {{pihua_start}}，{{title}}是{{time_query}}上課，{{pihua_end}}?
- **title**
    - {{pihua_start}}，{{instructor}}{{teacher}}{{course_query}}{{question}}，{{pihua_end}}?
    - {{pihua_start}}，{{instructor}}{{course_query}}{{question}}，{{pihua_end}}?

4. Backend Tables

For convenience, we now only create two general purpose tables for aforementioned four intents. (some attributes' values will be filled later)

- **Course** (raw_type: html-format (https://github.com/yvchen/CourseBot2/blob/master/crawler/md-based_course-table_example.html))

Semester	Serial_no	Designated_for	Curriculum_no	Class_no	Tit
104-1	73565	網媒所	CSIE5430		機器

- **Review** (raw_type: html-format (https://github.com/yvchen/CourseBot2/blob/master/crawler/md-based_review-table_example.html))

Title	Loading	Sweetness	Stars	Content
[評價] 105-1 臺灣史一 李文良	0	0	0	(是 / 否 / 其他條件) ：是 哪一學年 105-1 ψ 授課教師 (若為多人合授請寫 以方便收錄) 李文良 λ 開課系所與授 (是否為必修或通識課 / 內容是否與某些 歷史系大一必修 δ 課程大概內容 明末 ~開港前台灣史 Ω 私心推薦指數(以 ★★★★★ 滿天星 η 上課用 (影印講義或是指定教科書) 自編講 可於ceiba取得 μ 上課方式(投影片、 老師教學風格) 口述+一點點版 不過老師聲音頗輕柔，太累的話就 評分方式(給分甜嗎？是紮實分？) 好 每個星期都有指定閱讀， 紮實與否純看用功程度XD ρ 考題型式 三篇讀書心得，期末考申論題+名詞解 (是否注重出席率？如果為外系選 需先有什麼基礎較好嗎？老師個性？ 嚴禁遲到等...) 全簽，助教課要 正課好像是隨老師心情點名 Ψ 如果每個星期讀書的話其實會覺得正課 然後期末考都好簡單這樣XD 不過依照討論課的情況， 真的每星期讀的人好像不多 然後討論課刷存在感好像++++++很 讀書心得給分 我大概16個星期(18扣期中期末 認真讀了13個星期吧(圖書館大好Ow 有讀書的話期末考唯一的困擾就是

5. Backend Functions

Currently, we have only one function to handle the 4 supported intents

- **query_course(goal, slot)**

```
    def query_course(goal, slot):
    """Return list of Course objects
        Where goal in ['instructor', 'title', 'classroom', 'schedule']
        1. 自然語言處理在哪裡上課?
        query_course(goal='classroom', title='自然語言處理') -> '資105'
        2. 陳信希老師有開什麼課?
        query_course(goal='title', instructor='陳信希') -> ['專題研究', '
        3. 機器學習技法是在什麼時候上課?
        query_course(goal='schedule', title='機器學習技法') -> '二5,6'
    """
```

- ## Language Understanding

1. Training Data
Our training data is obtained from two sources:
- `TEMPLATE` : We generated training data with the templates described above. The generated setences are segmented with `jieba` . We give known slot values (e.g. course titles, instructor names) extremly large weights in a user dictionary so that `jieba` will not separate these values into multiple tokens. There are 4,219,218 instances in this dataset.
- `LOG` : We store the user inputs in a log file and manually correct the intent/slot labeling. We remove instances in which word segmentation errors make it impossible to label slot corrrectly, or the user intent cannot be clearly identified. After filtering, there are 182 instances in this dataset.

We extend the `LOG` dataset by plugging in different slot values. For example, we can use the following **source sentence**

```
陳縕儂 老師 開 了 什麼 課 ?
```

to generate the following data:

```
李宏毅 老師 開 了 什麼 課 ?
林軒田 老師 開 了 什麼 課 ?
...
```

We make 1000 copies for each original user input in `LOG` and randomly sample a course for each copy to fill the slots. The extended datset is called `E-LOG` .
After some experiments we found out that although there are huge amount of data in `TEMPLATE` , the model trained on it tend to overfit to the pattern of generating data and perform badly when processing real user input. Therefore, we extract a subset from `TEMPLATE` and `LOG-E` respectively to form the training set. The statistics are shown in the below table:

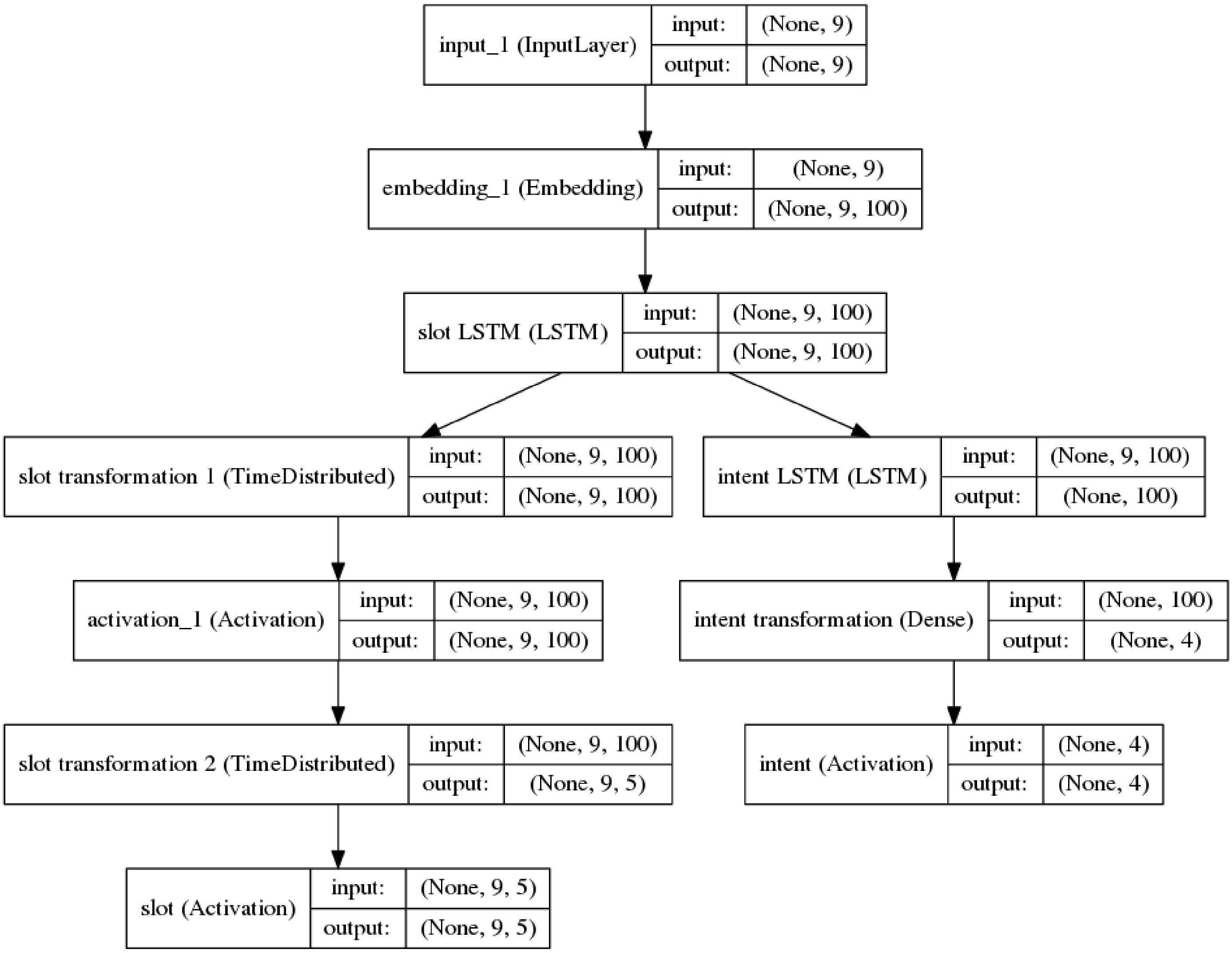
Source	#instances
`TEMPLATE`	1000
`E-LOG`	65,000
(total)	66,000

2. LU Model

We use LSTM to perform language understanding. Intent classification and slot filling are learned jointly. There are two LSTM layers in our model. The first LSTM reads the sequence of tokens in the input sentence and predict the BIO-encoding label of each token. The hidden state of each time step of first LSTM is also fed into the second LSTM for intent classification. The last hidden state of the second LSTM is used to determine the intent.

We use `Keras` to implement the LU model. 10% of the training data is utilized for validation. The training parameters and the configuration are shown below.

Loss function	Optimizer	Initial learning rate	Activation function	Weight of intent loss	Wei of s los
`categorical_crossentropy`	Adam	0.001	ReLU	0.8	0.:



In the above figure, the embedding layer size is set to 100. In fact we have experimented with different embedding size and the results will be shown later. The embeddings are randomly initialzed and will be updated throughtout the training process.

3. Performance Testing

We use "real" user input from `LOG` to test the performance. **No sentence in the test set is the source sentence of any intance in the training set.** By doing so, we can measure how well the model generalizes to new input. There are 117 test inctances. The results are shown in the below table.

Embedding layer size	**Intent (accuracy)**	**Slot-when (precision / recall)**	**Slot-title (precision / recall)**	**Slot-instructor (precision / recall)**
100	**0.872**	1.000 / 0.500	1.000 / 0.821	1.000 / 0.927
300	0.846	1.000 / 0.625	**1.000 / 0.859**	**1.000 / 0.976**
500	0.829	**1.000 / 0.750**	1.000 / 0.833	**1.000 / 0.976**

As can be seen, the model performs well in terms of precision, but it sometimes miss some slots. According to analysis of the model output, the prediction of intent is also affected when there are some errors regarding slot filling.
When increase embedding layer size from 100 to 300, there are some improvements on slot filling but the performance of intent classification is harmed. There is only slight improvement on one of the slots when moving to 500-dim. We finally choose the 100-dim model.

4. Examples and Predicted Semantic Frames

```
{
    'tokens': ['智慧對話機器人', '在', '哪間', '教室', '上', '?'],
    'labels': ['B_title', 'O', 'O', 'O', 'O', 'O'],
    'slot': {'title': '智慧對話機器人'},
    'intent': 'classroom'
}
```

```
{
    'tokens': ['離散', '是', '誰', '上', '課', '?'],
    'labels': ['B_title', 'O', 'O', 'O', 'O', 'O'],
    'slot': {'title': '離散'},
    'intent': 'instructor'
}
```

```
{
    'tokens': ['林軒田', '老師', '的', '機器學習', '什麼', '時候', '上', '課', '?'],
    'labels': ['B_instructor', 'O', 'O', 'B_title', 'O', 'O', 'O', 'O', 'O'],
    'slot': {'title': '機器學習', 'instructor': '林軒田'},
    'intent': 'schedule'
}
```

```
{
    'tokens': ['李宏毅', '老師', '這學期', '開', '了', '哪些', '課', '?'],
    'labels': ['B_instructor', 'O', 'O', 'O', 'O', 'O', 'O', 'O'],
    'slot': {'instructor': '李宏毅'},
    'intent': 'title'
}
```