110-2
Natural Language Processing

Final Project

TA: Kuei-Chun Kao

Task introduction (Pun Location)

- For each context, the system must identify which word is the pun
- Word sense disambiguation
- Homographic puns dataset
- What is homographic pun (語意雙關語)?
  - A homographic pun plays on words that are spelled the same way but have a double meaning. Because these puns rely on spelling, they are visual and must be read to be understood. Here is an example of a homographic pun that transposes the word “flies”: “Time flies like an arrow; fruit flies like a banana.”

Dataset

- Each context contains one pun
- Each pun (and its latent target) contains exactly one content word (i.e., a noun, verb, adjective, or adverb).
- Dataset format: xml format (need to parse by yourself)
- Example: sample.xml
- Corpus
  - Text (+)
    - Word (+)

Outputs

- Each line consists of two fields separated by horizontal whitespace (a single tab or space character). The first field is the ID of a text from the XML file. The second field is the ID of the one word in that text which is a pun.
- Kaggle Link: https://www.kaggle.com/t/0f854e5efb4e48f3a8369326cfe73a39
- Displayed name: <student_ID>
- Submission format: .csv file (You can also see from sample_submission.csv)
- Evaluation metric: F1 score

baseline.csv		0.70625

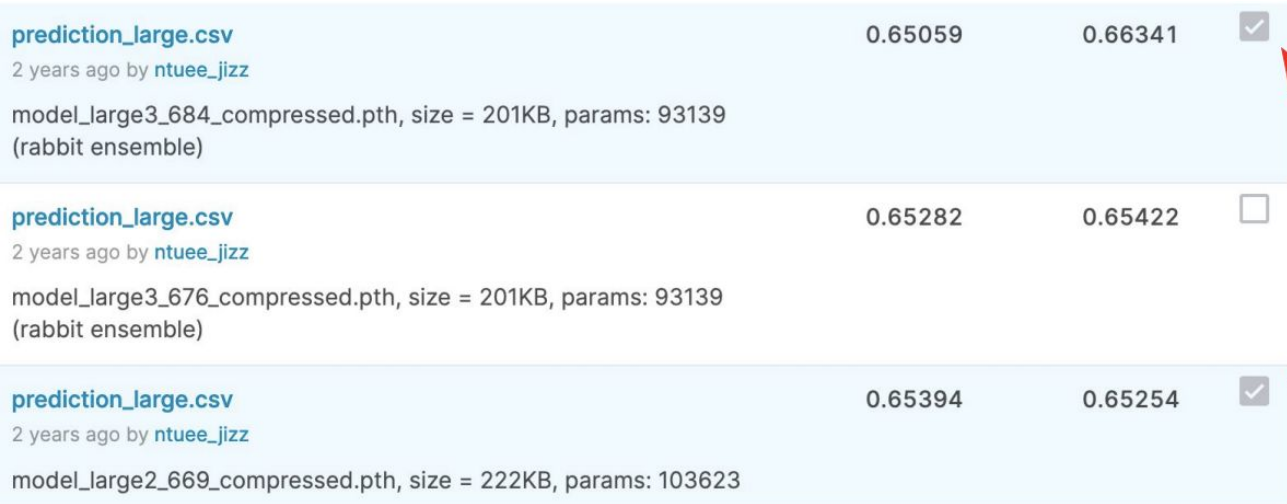
Example

```
<text id="t_1">
  <word id="t_1_0">I</word>
  <word id="t_1_1">used</word>
  <word id="t_1_2">to</word>
  <word id="t_1_3">be</word>
  <word id="t_1_4">a</word>
  <word id="t_1_5">banker</word>
  <word id="t_1_6">but</word>
  <word id="t_1_7">I</word>
  <word id="t_1_8">lost</word>
  <word id="t_1_9">interest</word>
  <word id="t_1_10">.</word>
</text>
```

```
t_1 t_1_9
```

Kaggle submission

- You may submit up to 5 results each day (UTC).
- Up to 2 submissions will be considered for the private leaderboard



remember to select **2** results for your final scores before the competition ends!

Reference (You can follow some ideas here)

- https://web.stanford.edu/~jurafsky/slp3/18.pdf
- Paper: A Unified Model for Word Sense Representation and Disambiguation (EMNLP-2014)

Requirements

- Python only
- No plagiarism!
- At the top of your Source code

```
#Author: Kuei-Chun Kao

#Student ID: 1234567

#HW ID: final_project

#Due Date: 01/30/2020
```

Submission

- Deadline
  - Submit Zip to E3 before 6/6 11:59 PM
  - No Late Submission, thanks!
- Format
  - Source code: final_<StudentID>.py (py only)
  - Report file: final_<StudentID>.pdf (pdf only)
  - Make sure the .py file contains the correct execution results and formats.
  - If can’t compile correctly, no score for you
  - Zip file: final_<StudentID>.zip (zip only)
- Any question can ask me on E3, answer your question ASAP

Grading policy

- Ranking score in Kaggle Leaderboard (40%)
- Report (40%)
- Final presentation (20%)
- I can only see your last submission.
- Do not submit your model or dataset.
- If your code is not reasonable, your final grade will be multiplied by 0.8!
- You should NOT modify your prediction files manually.
- Do NOT share codes or prediction files with any living creatures.
- Do NOT use any approaches to submit your results more than 5 times a day.

Ranking score in Kaggle Leaderboard (40%)

- Public leaderboard (20%): Your public leaderboard score > baseline, you can get 20% of this part; Otherwise, you can only get 10% of this part.
- Private leaderboard (20%): Your private leaderboard score * 20%
- This part score = public leaderboard + private leaderboard

Report (40%)

1. Your model design and concept (8%)
2. What kind of word sense representation used and experimented in your model (8%)
3. What problem did you face during the homework and how you solved (8%)
4. Error Analysis and Discussion (8%)
5. Compare and implement unsupervised method and supervised method (8%)

Final project presentation (20%)

- Date: 6/9(Thu)、6/16(Thu) [in class]
- 15 minutes per group
- Please introduce your motivation, methods, experimental results, discussion, and conclusion
- Presentation grading policy：
    - Creativity (30%)
    - Implementation (30%)
    - Findings or Discussion (20%)
    - Presentation (20%)
- Remember to submit the slides before your presentation

Bonus

- If your ranking is top 3 in class, you can get 3 points bonus in this hw final score!