Project 2: Chinese Discourse Relation Recognition

NATURAL LANGUAGE PROCESSING, NTU CSIE, SPRING 2017

授課教師: 陳信希 教授

助教: 顏安孜、薛祐婷

{d04922005, r04922029}@ntu.edu.tw

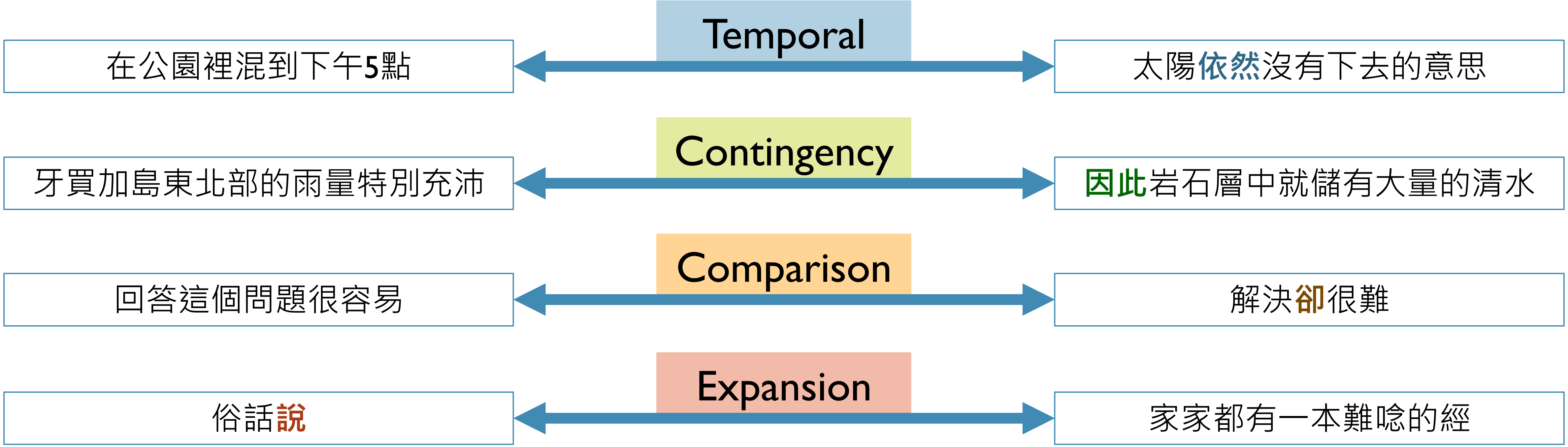
Date	Schedule
2017/05/18 (Thu.)	Project 2 release
2017/06/16 (Fri.)	Kaggle submission due
2017/06/18 (Sun.)	Report submission due
2017/06/22 (Thu.)	Final exam

OUTLINE

- Task Description
- Data
- Submission & Evaluation
- Grading & Rules
- Report
- Hints

TASK DESCRIPTION

- Recognize the **discourse relation** between a pair of clauses in a Chinese sentence



在公園裡混到下午**5**點 — Temporal — 太陽**依然**沒有下去的意思

牙買加島東北部的雨量特別充沛 — Contingency — **因此**岩石層中就儲有大量的清水

回答這個問題很容易 — Comparison — 解決**卻**很難

俗話**說** — Expansion — 家家都有一本難唸的經

DATA

- List of data files

1. `train.csv`: training pairs of clauses with discourse relation label
2. `test.csv`: test pairs of clauses
3. `sample_submission.csv`

DATA

- **train.csv**
  - 6,638 training instances
  - CSV columns
    1. **Id**
    2. **Clause1**
    3. **Clause2**
    4. **Relation**

```
Id,Clause1,Clause2,Relation
...
206,臺灣官方近來意識到都會公園的重要性,也積極擴展公園綠地的範圍。,Contingency
207,可以說,艙外服就是一個穿在身上的小型飛船。,Expansion
...
```

- **Relation** is one of `{Temporal, Contingency, Comparison, Expansion}`

DATA

- **test.csv**
  - 1,000 test instances (questions)
  - Format: same as `train.csv` but without the **Relation** column
    - In your submission, use the **Id** column to specify which question you are answering

```
Id,Clause1,Clause2
...
6659,為此,特向您表示衷心的感謝。
6660,不僅燈光更為醒目,行車安全也得到了提高。
...
```

- **sample_submission.csv**
  - CSV Columns:
    1. **Id**: question id in `test.csv`
    2. **Relation**: the most suitable discourse relation between the pair of clauses

```
Id,Relation
...
6659,Unknown
6660,Unknown
...
```

- You must replace the “Unknown” in the **Relation** column with your answers!

SUBMISSION & EVALUATION

- Submit your answers to the Kaggle platform
  - Join link: https://kaggle.com/join/ntunlp2017project2
    - Email to TA if you have problem joining the competition
  - A group may submit a maximum of 5 entries per day / select up to 2 final submissions for grading
  - Deadline: **2017/06/16 (Fri.) 23:59**
- Evaluation: $\text{accuracy} = \dfrac{\#\ \text{correctly answered questions}}{\#\ \text{test questions}}$
  - Public score (50% test data): will be updated in real time during the project
  - Private score (50% test data): will be announced after the deadline

GRADING & RULES

- Grading
  - Performance: 40%
    - Grade relatively
    - Mainly based on **private** score
  - Report: 60%

  - Each group will be treated equally **regardless of # members**

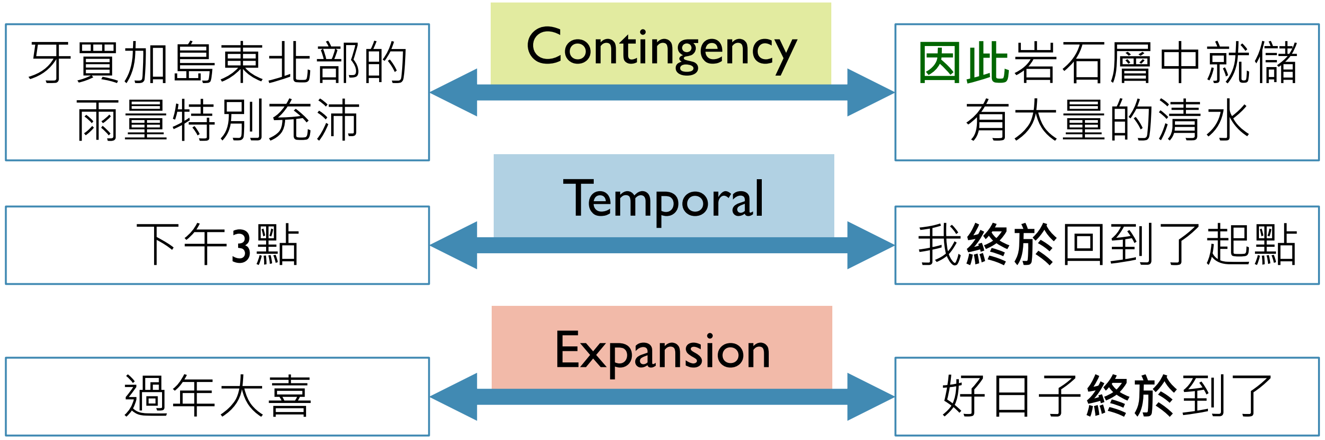
- Rules
  - You can
    - Use any toolkit/library
    - Use external text corpora
  - You **CANNOT**
    - Find the answer somewhere on the web
    - Manually answer the questions
  - **Contact the TAs if your are not sure!**

REPORT

- Report
  - Language: Chinese or English (Be clear in meaning!)
  - Pages: **no more than 6** (with readable font size)
  - Must include:
    - Name and student ID of each group member
    - Kaggle team name & email address for registering Kaggle
    - Agree to share your report in CEIBA 作業觀摩? (Yes/No)
    - Methodology, Experiments, Discussions
  - Deadline: **2017/06/18 (Sun.) 23:59**
    - Upload to CEIBA **in PDF format**, one copy per group

HINTS

- Discourse markers
  - Strong indicator of the discourse relation
  - But they can be **ambiguous**!



牙買加島東北部的雨量特別充沛

Contingency

因此岩石層中就儲有大量的清水

下午**3**點

Temporal

我**終於**回到了起點

過年大喜

Expansion

好日子**終於**到了

- Polarity
  - Different discourse relations, different tendencies
    - Comparison: positive-negative / negative-positive / …
    - Expansion: positive-positive / negative-negative / …
  - You can use the resources provided in Project 1
- You are encouraged to try other methods (not limited to those taught in class)

QUESTIONS?