Obesity
Status Detection

YM Group 3 :
蔡雯翔　李倍伊　丁玉芝　劉旭祐

01.

BACKGROUND

I2b2 obesity challenge

i2b2

Informatics for Integrating Biology and the Bedside

HARVARD MEDICAL SCHOOL | BLAVATNIK INSTITUTE BIOMEDICAL INFORMATICS

i2b2 (now n2c2)

A passionate advocate for the potential of existing clinical records to yield insights that directly impact healthcare improvement.

- **2006 - Deidentification & Smoking**
  - Evaluating the state-of-the-art in automatic de-identification
  - Identifying patient smoking status from medical discharge records
- **2008 - Obesity**
  - Recognizing Obesity and Co-morbidities in Sparse Data
- **2009 - Medication**
  - Extracting Medication Information from Clinical Text
  - Community Annotation Experiment for Ground Truth Gen
  i2b2 Medication Challenge
- **2010 - Relations**
  - 2010 i2b2/VA Challenge on Concepts, Assertions, and Re
  Clinical Text
- **2011 - Coreference**
  - Evaluating the state of the art in coreference resolution for electronic medical records
- **2012 - Temporal Relations**
  - Evaluating temporal relations in clinical text: 2012 i2b2 Challenge
  - Annotating temporal information in clinical narratives
- **2014 - Deidentification & Heart Disease**
  - Creation of a new longitudinal corpus of clinical narratives
  - Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1
  - Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus
- **2018 (Track 1) - Clinical Trial Cohort Selection**
  - Cohort selection for clinical trials: n2c2 2018 shared task track 1
- **2018 (Track 2) - Adverse Drug Events and Medication Extraction**
  - 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records

- **2008 - Obesity**
  - Recognizing Obesity and Co-morbidities in Sparse Data

TASKS

Design an **analysis flow** for obesity status classifiers according to textual judgment (presence of obesity or unmentioned).

Train_Textual

Training data based on textual judgement

– Textual judgement: 200 cases obesity vs. 200 cases unmentioned.

Test_Intuitive

Testing data based on intuitive judgement

– Intuitive judgement: 200 cases obesity vs. 200 cases absence

Validation

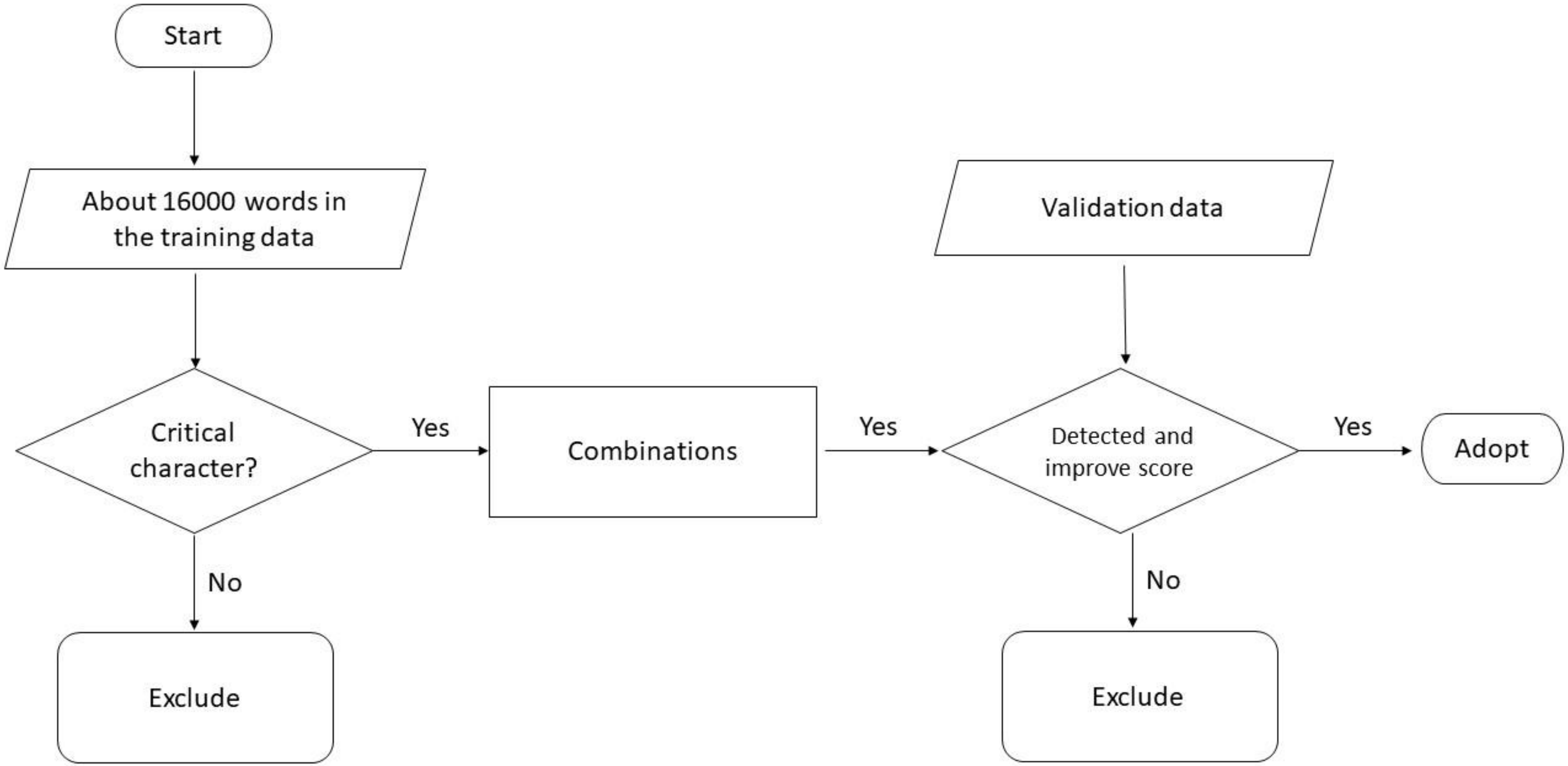
Validation data (50 cases) based on textual judgement

02.

METHODS

An implementation of bag-of-words model

PROCESS CHART



Import and get the characters in the medical record of training data

- unique
  - A function to combine the repeated characters
- Acommon: 16476

```
%% read medical records and convert them to characters
fl= dir('*.txt');
n = length(fl);

% read files and comebine the same character
lastrow = 0;
for j = 1:n
   symbolicseq = fileread(fl(j).name);
   text = text_preprocessing(symbolicseq,0);
   tra_text = text.';                              % transpose text
   uniq_text = unique(tra_text);                   % remove repetitions
   row = length(uniq_text);                        % get number of character

   for j = 1:row
       common(1,j+lastrow) = uniq_text(1,j);
   end

   lastrow = row+lastrow;
end
Acommon = unique(common);
```

Count the number of occurrences of each character in the obese and unmentioned groups

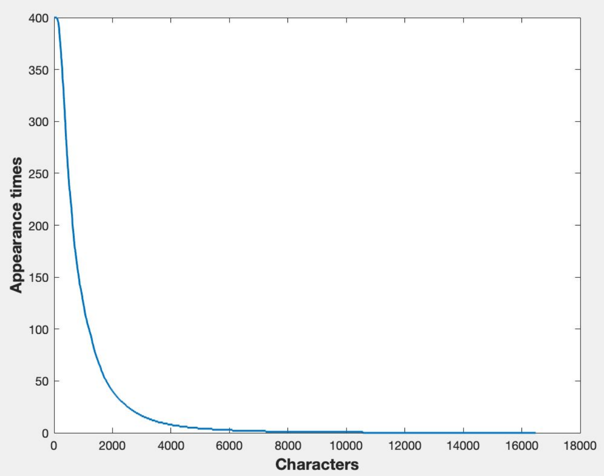
- Hypothesis: discriminative words only can be detected in the obesity group

```
for l = 1:2   % 1:capital; 2:lower
    if l == 2
        Acommon = lower(Acommon);
    end

    for j = 1:n
        mr = fileread(fl(j).name);
        find(j,:) = cellfun(@(s) ~isempty(strfind(mr, s)), Acommon);
        disp([fl(j).name, ' finish!'])
    end

    frequency = sum(find);
    obesity_fre(1,:) = Acommon;
    obesity_fre(2,:) = num2cell(frequency);
    obesity_tra = obesity_fre.';
    obesity_sort = sortrows(obesity_tra,2,'descend');
```

	1	2
1	'a'	400
2	'ab'	400
3	'ad'	400
4	'ar'	400
5	'as'	400
6	'at'	400
7	'ate'	400
8	'b'	400
9	'c'	400
10	'ch'	400
11	'co'	400
12	'ct'	400
13	'd'	400
14	'di'	400
15	'dm'	400
16	'dmi'	400
17	'e'	400



Compute the proportion of the character occurrences in two groups

1. Number of character
2. Number of occurrences in umentioned data
3. Number of occurrences in obesity data
4. The difference between obesity and unmentioned group
5. Percentage of occurrences in obesity to all cases

```
m = length(Acommon);

for j = 1:m
      score(1,j) = j;
      score(2,j) = sum(find(1:200,j));
      score(3,j) = sum(find(201:400,j));
      score(4,j) = score(3,j)-score(2,j);
      score(5,j) = score(3,j)/(score(2,j)+score(3,j));
end
```

1	2	3	4	5
5578	0	5	5	1
6857	0	5	5	1
7536	0	4	4	1
7583	0	5	5	1
7585	0	4	4	1
7611	0	4	4	1
8193	0	5	5	1
8524	1	10	9	0.9091
8724	0	5	5	1
8844	0	6	6	1
9079	1	9	8	0.9000
9326	0	5	5	1
10058	0	8	8	1
10080	1	9	8	0.9000
10493	0	4	4	1
10764	0	4	4	1
11290	1	10	9	0.9091
11609	0	6	6	1
12135	0	4	4	1

Exclude non-critical character

- Occurrences proportion lower than 0.9
- Less than 4 times

- WORDS : 36
- words : 74

```
    score = score.';
    del = score(:,5)<0.9| (score(:,4)<4 & score(:,5)>=1) | isnan(score(:,5));
    score(del,:)=[];

    words_result = Acommon.';
    words_result(del,:)=[];

    if l == 1
        WORDS = words_result;
        WORDS_score = score;
    else
        words = words_result;
        words_score = score;
    end
end
```

	1
1	APNEA
2	BILL
3	CALL
4	CHRONIC
5	CRUZ
6	DIAGNOSTIC
7	DOA
8	EDWARDO
9	FOAT
10	FROM
11	FUNCTION
12	IGNACIO
13	JONATHAN
14	JUST
15	JUSTIN
16	KARL
17	LHC
18	LW
19	MARK
20	MCH
21	MICONAZOLE
22	MORPHINE
23	OBESITY
24	OBSTRUCTIVE

	1
1	acquired
2	antacids
3	atrovent
4	bloating
5	brisk
6	bubble
7	cefpodoxime
8	collar
9	commands
10	community
11	cpa
12	cpap
13	cultured
14	cyanotic
15	declining
16	decompress
17	dentition
18	discs
19	draining
20	exhibit
21	expiratory
22	exudative
23	feelings
24	flagy

Find the best combination in testing data

- WORDS + words = 110
- The number of combinations
  - Two : 5,995
  - Three : 215,820
  - Four : 5,773,185
  - Five : 122,391,522

	1	2
1	obese	obesity

	1	2	3
1	dentition	obese	obesity

	1	2	3	4
1	dentition	obese	obesity	pod

	1	2	3	4	5
1	CHRONIC	dentition	obese	obesity	pod

	Combinations	Precision	Accuracy	Recall	F1-score
	1	2	3	4	5
1	1x2 cell	0.9060	0.8025	0.6750	0.7736
2	1x2 cell	0.9167	0.7500	0.5500	0.6875
3	1x2 cell	0.9279	0.7375	0.5150	0.6624

	1	2	3	4	5
1	1x3 cell	0.9085	0.8125	0.6950	0.7875
2	1x3 cell	0.9026	0.8100	0.6950	0.7853
3	1x3 cell	0.9079	0.8100	0.6900	0.7841

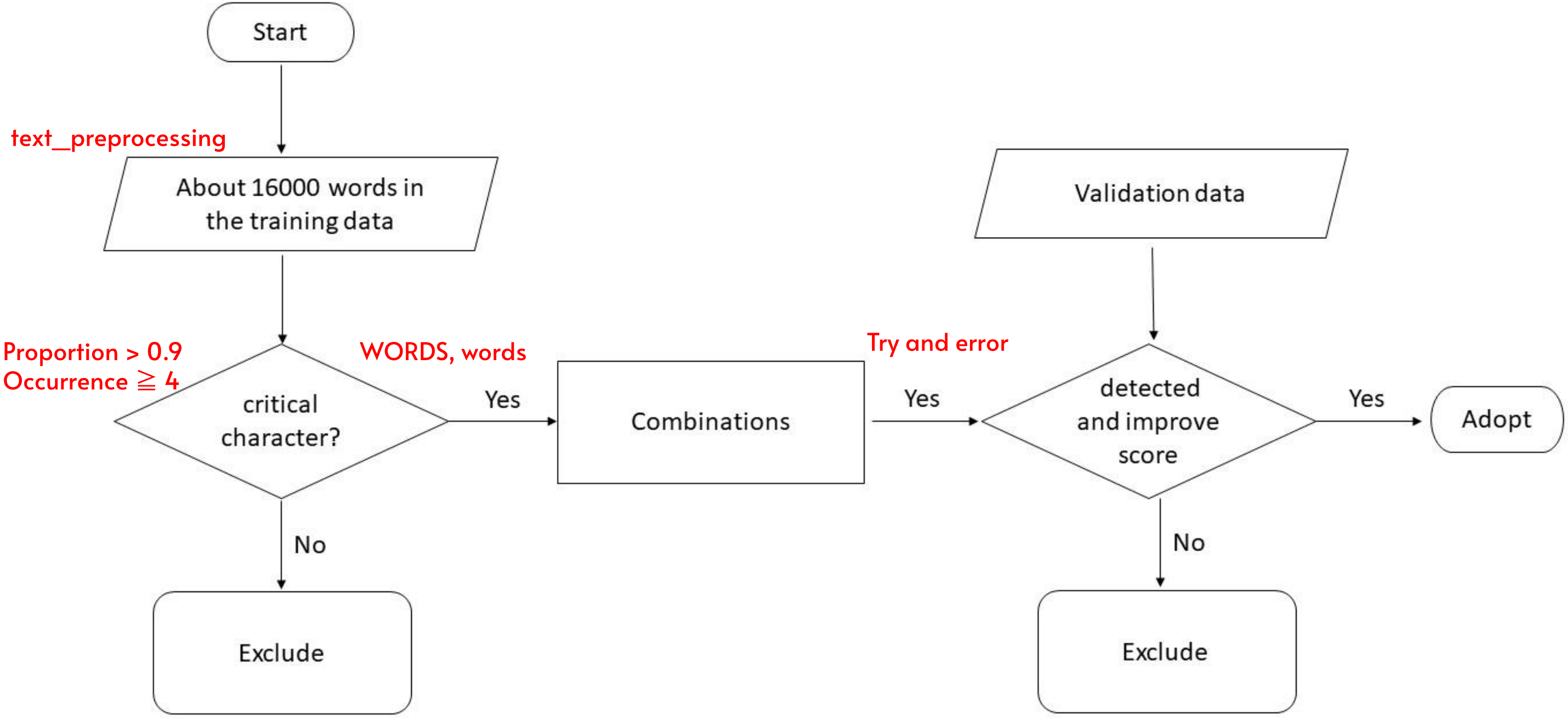
	1	2	3	4	5
1	1x4 cell	0.9051	0.8200	0.7150	0.7989
2	1x4 cell	0.9103	0.8200	0.7100	0.7978
3	1x4 cell	0.9045	0.8175	0.7100	0.7955

	1	2	3	4	5
1	1x5 cell	0.9012	0.8250	0.7300	0.8066
2	1x5 cell	0.9063	0.8250	0.7250	0.8056
3	1x5 cell	0.9063	0.8250	0.7250	0.8056

Pick the top characters

- Basic
  - obese, obesity
- Analysis result
  - CHRONIC (慢性的), morbid (病態的), hypokinetic (運動不足), dentition…
- High correlation words
  - Cerebrovascular (腦血管), Obstructive, OSA (Obstructive sleep apnea)

PROCESS CHART



GitHub

https://github.com/WenXiangTsai/DM_CasePresentation

WenXiangTsai Update README.md ...	2 hours ago	13
README.md	Update README.md	2 hours ago
obeseidx.m	Update obeseidx.m	15 hours ago

03.
VALIDATION RESULTS

F1- score on validation sets

Validation result

Try different combination of key characters in the verification data

- Try and error

Words	F-score
obesity','obese'	0.51428
obesity','obese','CHRONIC'	0.57142
obesity','obese','dentition','pod'	0.54285
obesity','obese','dentition'	0.54285
obesity','obese','dentition','CHRONIC'	0.60000
obesity','obese','dentition','isosorbide'	0.54285
obesity','obese','CHRONIC','dentition','cultured'	0.60000
obesity','obese','CHRONIC','dentition','community'	0.57142
obesity','obese','CHRONIC','dentition','cerebrovascular'	0.62857
obesity','obese','CHRONIC','dentition','angiography'	0.57142
obesity','obese','CHRONIC','dentition','Obesity'	0.60000
obesity','obese','CHRONIC','dentition','Obese'	0.57142
obesity','obese','CHRONIC','dentition','nebulizer'	0.54285
obesity','obese','CHRONIC','dentition','CPAP'	0.60000
obesity','obese','CHRONIC','dentition','ADVAIR'	0.62857
obesity','obese','CHRONIC','dentition','ADVAIR','OSA'	0.60000
obesity','obese','CHRONIC','dentition','ADVAIR','cerebrovascular'	0.65714

04.

DISCUSSION

Limitation

- Bag-of-words model
  - Lack of analysis of numbers and context
- The analysis method is too specific to this test data
- Lack of misspelling detection
- We do not identify specific zones e.g. discharge summaries, past medical history etc., which might lead to more precise analysis

Reference

Website:

- n2c2 NLP Research Data Sets
- 肥胖症-維基百科
- Obesity and overweight - WHO
- Recognizing Obesity and Comorbidities in Sparse Data

Photo:

- Study Finds Obesity Itself Raises Risk of Diabetes and Cardiovascular Disease

Slide template:

- Healthcare Center Website - slidesgo

Contributions

蔡雯翔: 討論、投影片製作、資料分析、github readme撰寫、報告

李倍伊: 討論、投影片製作設計

丁玉芝: 討論、臨床經驗分享

劉旭祐: 討論、投影片製作、github readme撰寫

Thank you for your time!