Natural Language Processing Term Project

NTU CSIE, Spring 2015

Term Project Goal

- Working on a research issues
  - Be familiar with common NLP methodology
  - Make flexible use of common NLP tools
  - Understand the algorithms learned in class better
  - Implement some algorithms on your own
  - Learn how to do research
  - Read relevant papers
  - Brainstorming
  - Teamworking
  - ...Many other things

Motivation

- Chinese as a foreign language
  - Some mistakes are made by learners

Motivation

- Chinese as a foreign language
  - Some mistakes are made by learners

就不能实现的幸福了

Motivation

We want to know

- Whether the sentence contains error or not
- Where is the error

Teamwork

- 2~3 people (If you want to have one-person team, you are supposed to be as good as one team)
- Please fill in team members' names by 4/2
  - http://goo.gl/5Hbb5I
- If you can't find other team members, you can try recruiting your team members on the forum of CEIBA or anywhere else.

Phase I: Redundant Error Detection

- Training data
  - Sentences labeled with
    - 1: contains redundant word error
    - 0: does not have redundant word error
- Testing data
  - 200 sentences
  - Once testing data is released, you have 1 week to submit the results

Redundant Word Error

就不能实现的幸福了

Phase I: Training Data Format

- UTF-8 no BOM, lines ended with LF
- Filename: p1.train.txt
- Each line has:
  - ID <TAB> 1/0 <TAB> SENTENCE
- Example:
  - p1train-x	1	就不能实现的幸福了
  - p1train-y	0	这不是传统宗教势力的“反扑”

Phase I: Testing Data Format

- Filename: p1.test.txt
- Each line has:
    - ID <TAB> SENTENCE
- Example:
    - p1test-x   就不能实现的幸福了

Phase I: Result Format

- Each line has:
  - ID <TAB> 1/0
    - 1: contains redundant word error
    - 0: does not have redundant word error
- Example: http://goo.gl/G09rdz
  - p1test-x	1

Phase I: Submission Requirement

- Submit
  - The result for testing data
  - Your system source code which can successfully run in case that TA may check them
- Zip them into onefile, Phase1_groupID.zip and Submit from CEIBA
  - /p1.result.txt
  - /source/...
- Each group only needs to submit once

Phase II: Locate Redundant Word

- Training data
  - Sentence pairs
    - Sentence with redundant word
    - Corrected sentence
- Testing data
  - 200 sentences with redundant word
  - Once testing data is released, you have 1 week to submit the results

Phase II: Training Data Format

- UTF-8 no BOM, lines ended with LF
- Filename: p2.train.txt
- Each line has:
  - ID <TAB> SENTENCE-E <TAB> SENTENCE-C
- Example:
  - p2train-x 就不能实现的幸福了 就不能实现幸福了

Phase II: Testing Data Format

- Filename: p2.test.txt
- Each line has:
  - ID <TAB> SENTENCE
- Example:
  - p2test-x	就不能实现的的幸福了

Phase II: Result Format

- Each line has:
  - ID <TAB> START <TAB> END
- Example:
  - p2test-x 6 7
- Explanation:
  - 就不能实现的的幸福了
  - 1 2 3 4 5 6 7 8 9 10
  - The 6th & 7th character is redundant, so start and end are [6, 7]
  - A character may be Chinese character, symbol, number, English letter, etc. The testing data won’t contain spaces.

Phase II: Testing & Result Format

- We will release additional data that have the same format as testing & result for your reference
    - p2.val.txt
    - p2.val.ans.txt

Phase II: Submission Requirement

- Submit the result for testing data
- Report for both Phase I & II (at most SIX A4 pages with readable font sizes in pdf)
- Your system source code which can successfully run in case that TA may check them
- Zip them into onefile, Phase2_groupID.zip and Submit from CEIBA
- Each group only needs to submit once

Phase II: Submission Requirement

- Zip them into onefile, Phase2_groupID.zip and Submit from CEIBA
  - /p2.result.txt
  - /Report.pdf
  - /source/...

Evaluation: Phase I

- F-1 score for erroneous sentence detection (i.e. the label 1)
- Example
    - Truth: 0, 0, 1, 1
    - Your result: 1, 0, 1, 1
        - Recall: 1
        - Prec: 0.66
        - F-1: 2 * 1 * 0.66 / (1+0.66)

Evaluation: Phase II

1. Number of completely correct cases
2. Macro-average F-1 scores for each case

- Example
    - Truth: 5 7
    - Your result: 6 10
        - Recall: 0.66, {6th, 7th} / {5th, 6th, 7th}
        - Prec: 0.4, {6th, 7th} / {6th, 7th, 8th, 9th, 10th}
    - Use averaged Recall & Prec to calculate F1

Report

Must cover both Phase I & Phase II

- Introduction
- Your observation
- Your idea & method
- Self evaluation
- Discussion
- …….
- At most SIX A4 pages with readable font sizes in pdf

A detailed and well structured report will get bonus grades

Scoring

- Performance on testing data, innovative (algorithm), good evaluation and experiment design, good idea, good discussion, good trying, good observation, homemade tools, good efficiency ……
- everything good may get good grades, so write them on the report
- If you failed in many test cases, you can analyze why your method didn’t work and write them on the report
- It is still ok if you can’t get good results (error detection is hard!), just write all the things you do in the report , you still can get enough grades for your effort paid

Important Dates - http://goo.gl/tNp0gr

2015/3/26	Term project announced, Phase 1 training data released.
2015/4/2, 23:59	Fill in the team sheet
2015/4/23	Mid-term Exam
2015/4/27	Phase 1 test data released.
2015/5/4, 23:59	Phase 1 due, submit your phrase 1 result and code.
2015/5/5	Phase 2 training data released.
2015/5/21	Phase 1 presentation.
2015/6/2	Phase 2 test data released.
2015/6/9, 23:59	Phase 2 due, submit your phrase 2 result, code, and phrase 1 & 2 report.
2015/6/25	Final Exam

Reference

You can use any tools and data sets to help you. (But you cannot use the ground truth, if you happen to know the source.)

- Toolkits
  - https://g0v.hackpad.com/aco0Hxp4IEz
  - https://github.com/josephmisiti/awesome-machine-learning
- Google Ngrams
  http://storage.googleapis.com/books/ngrams/books/datasetsv2.html
- Similar Data Sets
  - http://ir.itc.ntnu.edu.tw/lre/nlptea14cfl.htm
  - http://ir.itc.ntnu.edu.tw/lre/sighan7csc.html

Any Questions?

- Please ask now
- Or you can email to: nlp2015ta@gmail.com