Natural Language Processing and Text Mining: HW#2

By J. H. Wang

Apr. 10, 2023

Programming Exercise #2: NER

- Goal: Named Entity Recognition on open datasets

- Input: BTC NER dataset (to be detailed later)

- Output: Training a model to recognize the named entity types (to be detailed later)

Tasks and Data

- Tasks
  - Performing NER on Twitter data (as detailed in the following slides)

- Data: an open dataset available from GitHub

- You have to submit the result of NER in terms of the F1 score

Input Data

- Data:
  - [**Broad Twitter Corpus**] available from GitHub
  - Available at:
    - https://github.com/juand-r/entity-recognition-datasets
- Format:
  - 6 files in CoNLL format
  - Each line contains:
    token ner_tag
    - BIO or IOB format

IOB Format (Inside-Outside-Beginning) – An Example

- IOB format:
  - Alex I-PER
  - is O
  - going O
  - to O
  - Los I-LOC
  - Angeles I-LOC
  - in O
  - California I-LOC

- IOB2 format:
  - Alex B-PER
  - is O
  - going O
  - to O
  - Los B-LOC
  - Angeles I-LOC
  - in O
  - California B-LOC

Sections in the dataset

Section	Region	Collection period	Description	Annotators	Tweet count
A	UK	2012.01	General collection	Expert	1000
B	UK	2012.01-02	Non-directed tweets	Expert	2000
E	Global	2014.07	Related to MH17 disaster	Crowd & expert	200
F	Stratified	2009-2014	Twitterati	Crowd & expert	2000
G	Stratified	2011-2014	Mainstream news	Crowd & expert	2351
H	Non-UK	2014	General collection	Crowd & expert	2000

Tasks in this Homework

- To train a model to recognize the named entity types in English
  - The program could be written in any programming language
  - You can write your own models or call existing open source APIs in your program
  - Please specify the platform and compilation instructions in your documentation
- To output the result of NER in terms of the F1 score

Some example implementation

- Open source APIs or libraries
    - Nltk, SpaCy (in Python)
    - Stanford NER, OpenNLP (in Java)
    - HanLP, CKIP CoreNLP (for Chinese)
    - …
- Implementation methods:
    - CRF: Conditional Random Field
    - HMM: Hidden Markov Model
    - RNN, LSTM
    - BERT
    - …

Output Format

- recognition results
  - Precision
  - Recall
  - F-measure
  - Accuracy

Homework Submission

- Due: three weeks, May 1 , 2023 (Mon.)

- For programming exercises, please submit it online to iSchool+
  - Under the item [Assignments]\[HW#2]

- Please include program source codes and documents
  - specifying your team members and responsible parts in the homework
  - Indicating configuration and installation steps of necessary packages on the specified platform

References

- Broad Twitter Corpus: A Diverse Named Entity Recognition Resource. Leon Derczynski, Kalina Bontcheva, and Ian Roberts. Proceedings of COLING, pages 1169-1179, 2016.

Thanks for Your Attention!