README

This repository includes python source codes, Jupiter files, result analysis and dataset containing 12,344 patent records for the tasks Patent Classification. The program has employed a deep learning method BiLSTM [1] to classify patents based on Invention Titles, Abstract Context with word embeddings.

1.Library requirements

- tensorflow (>=2.0)
- keras(>=2.2)
- numpy (>=1.12)
- nltk(>=3.2.0)

2. Technical program

+ Data set: uspto.csv (the data is generated from patent xml files [2] using data_lean.py)
- uspto.csv contains 12,344 patent records with a structure.
  Col 0,1: Invention no. and Date
  Col 2: Invention title
  Col 3: Main category (A, B, C, D, E, F, G, H)
  Col 4: Sub-categories
  Col 5, 6, 7: Sub-categories (for further purposes)
  Col 8: Abstract context
  Col 9: Body context

+ Data preprocessing
- Assigning max length of a record (Invention Titles=20, Abstract Context=100)
- Splitting data into 85% for training and 15% for validation.
- Removing stopword.
- Transferring data sequences to tokens.
- Padding data tokens to the max length.

+ LSTM implementation
- Using tf.keras.Sequential model
- Adding an embedding layer expecting input vocab of size, and output embedding dimension of size 64.
- Using a Dense layer (9 units for Invention Titles, and 479 for Abstract Context) and Softmax activation.

+ Training
- Use epochs=15 to train model.
- Model could be fitted at optimized epochs.

+ Prediction
- Sample input text = “Apparatus and method for determining a physiological condition” as invention title.

Output: [[0.00051499 0.07950258 0.10365563 0.19542012 0.26217026 0.25165954 0.05093597 0.05289719 0.00324374]]

A

3. Demo

a) Execute a command as: **python lstm_uspto.py para1 para2** (for Word2Vec)

python lstm_glove_uspto.py para1 para2 (for GloVec)

+ para1:
    - cat1: main category
    - cat2: sub category
+ para2:
    - inv: using Invention Title for classification
    - abs: using Abstract Context for classification

Note: For GloVec, “glove.6B.50d.txt” is downloaded from https://nlp.stanford.edu/projects/glove/

b) Check in Jupiter files

4. References

[1] https://bulkdata.uspto.gov/data/patent/grant/redbook/fulltext/2020/

[2] Liu, P., Qiu, X., Huang, X. (2016). Recurrent Neural Network for Text Classification with Multi-Task Learning. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)