Coursework 1: Question classification COMP61332 Text Mining

Nhung Nguyen

Task definition and data

- Question classifier:
  - Input: a question
    - How many points make up a perfect fivepin bowling score ?
  - Output: one of N predefined classes
    - NUM:count, i.e., counting questions

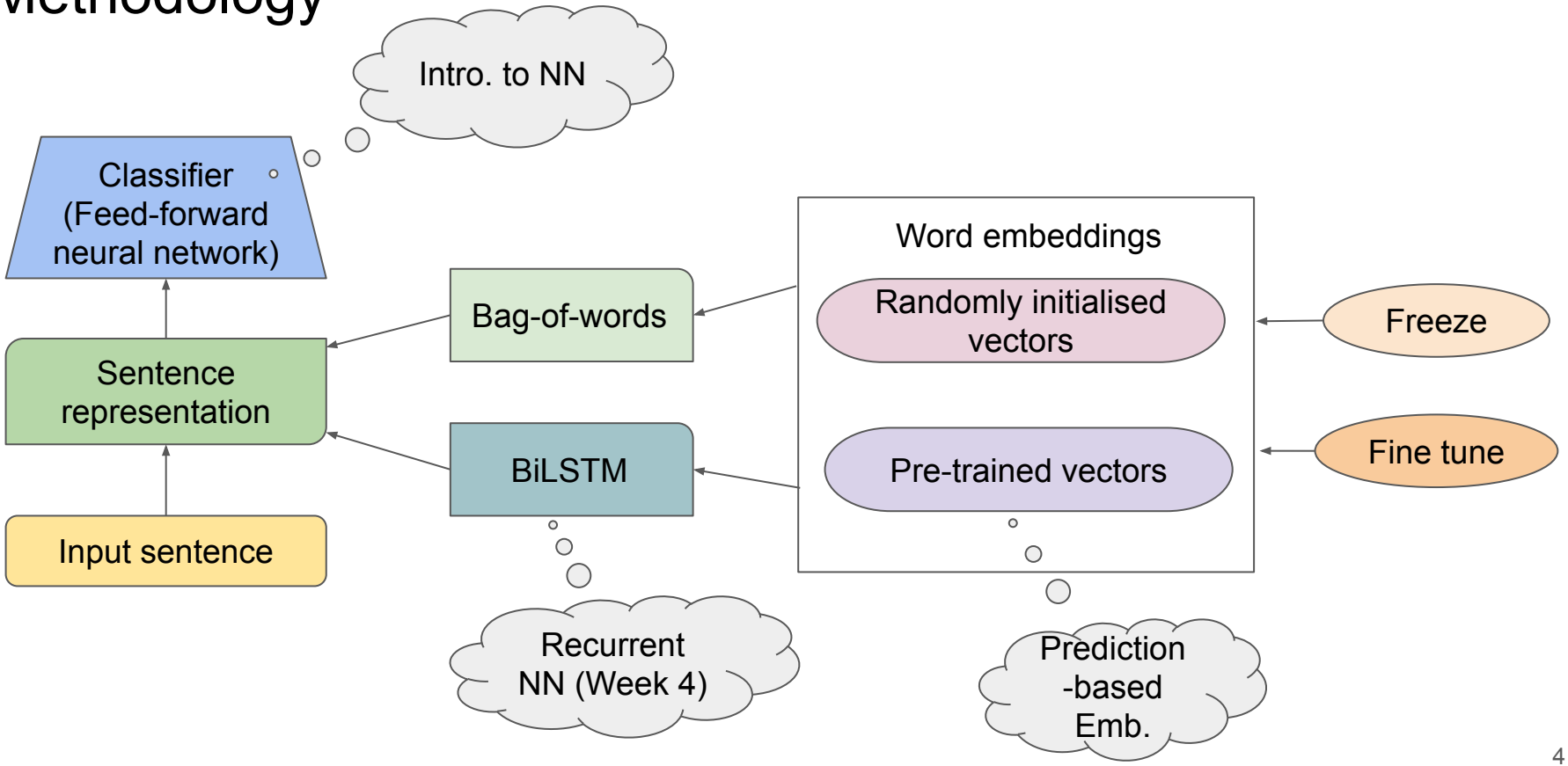
- Data: https://cogcomp.seas.upenn.edu/Data/QA/QC/
  - Training: Training set 5 (5500 questions)
  - Testing data: TREC 10 questions

Supervised learning framework

- Training stage:
  - Train the model on the training set (train)
  - Fine-tune/optimise the model on the development set (dev)
- Testing stage:
  - Test the model on the testing set (test)

- Most datasets have their splits with train/dev/test, but the aforementioned dataset does not
  - You **have to split the training set into 10 portions**. 9 portions are for training, and the other is for development.

Methodology



Deliverable 1 - Your implementation

- You can use any environment/operating system for your development, but TAs will use the school’s virtual machine to mark
- **Only** pytorch, numpy, and python3 standard libraries are allowed.
  - You don't need any off-the-shelf NLP libraries
  - Exceptionally: sklearn library for evaluation metrics, and other libraries for your interface.
- Please organise your source code as required

Deliverable 2 - A short paper

- Should be in the form of a research paper (2-3 pages excluding references)
- Should contain at least the following points:
  - Introduction/background
  - Your approach
  - Your experiments
    - Settings
    - Results
    - Analysis
  - Conclusion (if any)

Intended Learning Outcomes

- to develop deep learning-based sentence classifiers using word embeddings and BiLSTM
- to evaluate and analyse your sentence classifiers according to different settings
- to discuss your methods and results in the form of academic writing
- to practise teamwork skills

Deadline: Midnight of 12th March, 2021 (UK Time)
Good luck!