Assignment 2-Part 2
Overlap based WSD

CS626: Speech and Natural Language Processing and the Web

Problem statement

- Build an **overlap based** word sense disambiguation system using **word2vec** embeddings. Use the similarity between the word embeddings as a measure to compute overlap between sense bag and context bag
- Dataset: **SemCor 3.0 (Sense-tagged Corpus)**
- Input: **A sentence**
- Output**: WordNet sense ids for the words in the sentence**
- Create a document which reports the following
    - P, R, F1-scores
    - Compare the performance of your sense-tagger against Most Frequent Sense (MFS) baseline and WordNet (WN) 1st sense baseline
    - Perform detailed error analysis

Note:

- Use 5-fold cross-validation for reporting all scores. A helper code to generate word2vec embeddings for a given word has been provided in the next page
- Use nltk to download and access SemCor

Submission instructions

- The assignment is to be submitted in groups of 3 (Same group for every assignment and project)
- The submission link will be created on moodle to submit the assignment
- Only one person from the group with the lowest id is supposed to make the submission
- The name of the folder should be <id1_id2_id3>_Assignment2.zip
    - The uncompressed folder should contain code, readme and the slides used for presentation
    - The readme should contain details about the tools, versions, pre-requisites if any, and how to run the code

Deadline

- No-Hard deadline (Continuous Evaluation). Evaluation date will be announced soon

Generation of word2vec embeddings:

```
#Download pre-trained word2vec using the following command
!wget -c "https://s3.amazonaws.com/dl4j-distribution/GoogleNews-vectors-negative300.bin.gz"

#Import necessary modules and load word2vec
from gensim.models import KeyedVectors
model_w2v = KeyedVectors.load_word2vec_format('GoogleNews-vectors-negative300.bin.gz',
binary=True)

# Generate word2vec vectors for words (v represents the word2vec embedding for the word ‘language’)
v = model_w2v.wv[‘language’]
```