Assignment 2-Discussion <WSD>

Deepak Singh Baghel 203050005
Ankush Agarwal 203050007
Nilesh Kshirsagar 203059004
…

Problem Statement

- Given a sequence of words, produce the synset IDs (unique identification for a sense)
- Data SemCor
- First Technique to be used: HMM-Viterbi, Overlap based WSD
- 5-fold cross validation (for HMM based WSD)
- See if the baselines are crossed

First Baseline

- Most Frequent Sense (MFS)

- P: 0.60 R: 0.70 F-1: 0.66

Performance report of HMM-WSD

- Precision: 0.57, 0.58, 0.56, 0.57, 0.57
- Recall: 0.66, 0.66, 0.65, 0.65, 0.66
- F1-score: 0.58, 0.58, 0.57, 0.57, 0.58
- On 5 folds

Confusion Cases (for HMM)

- Actual : own.v.01 Predicted : have.v.01
  - Target word: has, had.
- Actual : think.v.01 Predicted : think.v.02
  - Target word: thought, think.
- Actual : yield.v.01 Predicted : give.v.01
  - Target word: given.
- Actual : foot.n.01 Predicted : foot.n.02
  - Target word: foot.
- Actual : receive.v.02 Predicted : get.v.01
  - Target word: got, getting.

Interpretation of confusion (error analysis: HMM)

- Words like ‘own’ and ‘have’, ‘yield’ and ‘give’, ‘receive’ and ‘get’ are used interchangeably by human hence led to confusion.
- More frequent senses are often getting predicted as compared to senses whose frequency in corpus is less.

Data Processing Info (Pre-processing: HMM)

- Stored count of sense and sense bigrams in a dictionary using sense as key and used it to calculate the transition prob.
- Similarly stored count of word sense pairs in a dictionary and used it to calculate the emission prob
- Use nltk.semcor.tagged_sents for tokenization.
- Lower casing

Performance report of Word Vector Based Overlap approach

- Precision : 0.71
- Recall : 0.495
- F1-score : 0.56

Confusion Cases (for Word Embedding and Overlap)

- Actual : antique.s.02 Predicted : outmode.v.01

  (Here ‘outmoded’ is used in adjective satellite sense but predicted sense is verb)
- Actual : evidence.n.01 Predicted : evidence.n.03

  (Here sense should be evidence.n.01 - but sentence included some words such as ‘investigation’ present in the sentence as well as sense definition of evidence.n.03)
- Actual : city.n.02 Predicted : city.n.01

  (Probably due to similarity in definition of two senses)

Confusion Cases (for Word Embedding and Overlap)

- Actual : junior_high_school.n.0 Predicted : junior.s.02

  (Here sentence was ‘junior or senior high school’ but actual sense was not taken into account for comparing with context because grouping of words was not done correctly)

- Actual : notag Predicted : inch.n.01

  (Similarly ‘in’ was predicted as ‘inch’, ‘or’ was predicted as ‘operating room’)

Interpretation of confusion (error analysis: WE-Overlap)

- For some words, such as ‘outmoded’, POS in which it is used is predicted wrongly
- Context may not have maximum similarity with correct sense in some cases
- Some function words were predicted to be abbreviations
- Grouping of words in dataset is not correct in some cases
- Senses having closely related definitions are often mispredicted

Data Processing Info (Pre-processing: Overlap based)

- Used gensim to access pre-trained word2vec embeddings.
- Stopword removal
- Used sense definition and example for comparing with context