Assignment 2-Discussion <WSD>

Ashwani Kumar Jha 20305R001
Lackarsu Siri Verma 180050052
Mikhil Gupta 203079016

7 September 2021

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)

- Precision: 0.5788
- Recall: 0.5097
- F1_score: 0.5415

Performance report of HMM-WSD

- Precision - 0.4054
- Recall - 0.4037
- F1-score - 0.4053
- On 1 fold for a limited size of test data (50 sentences)

Confusion Cases (for HMM)

It urged that the city take steps to remedy this problem

steps:

- Actual tag: ‘step.n.01’ : any maneuver made as part of progress toward a goal
- Predicted tag: ‘step.n.03’ : the act of changing location by raising the foot and setting it down

---

Nevertheless, we feel that in the future …

feel:

- Actual tag: ‘feel.v.02’ : come to believe on the basis of emotion, intuitions, or indefinite grounds
- Predicted tag: ‘feel.v.01’ : undergo an emotional sensation or be in a particular state of mind

Only a relative handful of such reports was received, …

only:

- Actual tag: ‘only.r.01’ : and nothing more
- Predicted tag: ‘only.r.02’ : without any others being included or involved

handful:

- Actual tag: ‘handful.n.01’ : a small number of amount
- Predicted tag: ‘handful.n.02’ : the quantity that can held in the hand

received:

- Actual tag: ‘receive.v.02’ : receive a specified treatment (abstract)
- Predicted tag: ‘receive.v.01’ : get something, come into possession of

Interpretation of confusion (error analysis: HMM)

- Misclassification of Grouped Words-In most of the cases HMM model was not able to correctly predict the sentence for group word.

  Eg- NE Fulton county Grand Jury
- Here we are only considering Bigram assumption,but actual sense depends on whole sentence context.
- In some cases, the POS tag of the word is predicted incorrectly, further adding to the error.

Data Processing Info (Pre-processing: HMM)

- Calculated the index of test and train data for each fold validation.
- Lowercase the sentence
- Added ‘^’ as first word/tag in each sentence
- Calculated frequency of emission_matrix[tag][word] and transition_matrix[prev_tag][tag] based on the training data
- Calculated probability of emission_matrix[tag][word] and transition_matrix[prev_tag][tag] based on the calculated frequency
- Created a dictionary to store tags for a particular word

Performance report of Word Vector Based Overlap approach

- Precision - 0.2736
- Recall - 0.2614
- F1-score - 0.2673

Confusion Cases (for Word Embedding and Overlap)

We have not analysed the difference between predicted and actual sense tags yet. However, one major part of error is associated to the fact that there is no word2vec embedding available for a group of words.

Interpretation of confusion (error analysis: WE-Overlap)

- <try giving reasons>

Data Processing Info (Pre-processing: WE-Overlap)

- The whole dataset is taken as the test data.
- For each sentence in the test data =>

  context_emb = sum of word vector of each word in input /# of words in input
- For each word in the sentence, there are a list of senses associated with it, for which sense embedding vectors are calculated.

  sense_emb = sum of word vector of each word in Gloss /# of words in Gloss
- Using the cosine similarity between sense embedding and context embedding for each sense of a word, the best sense is picked for a word in a given input sentence.
- This is repeated for all the sentences in the test data, and the predicted sense tags are recorded.
- To compare the predicted and actual tags, the detailed lemma tag of words in the dataset have been converted to synset tags.