Assignment-Discussion
HMM-Viterbi

Deepak Singh Baghel 203050005
Ankush Agarwal 203050007
Nilesh Kshirsagar 203059004

Problem Statement

- Given a sequence of words, produce the POS tag sequence
- Technique to be used: HMM-Viterbi
- Use Universal Tag Set (12 in number)
- 5-fold cross validation
- `tags: {'ADP', 'CONJ', 'PRON', 'DET', 'PRT', 'VERB', 'NOUN', 'NUM', 'X', 'ADJ', '.', 'ADV'}`

Overall performance

- Precision : 93.55 %
- Recall : 93.50 %
- F-score (3 values)
  - F1-score: 93.46 %
  - F0.5-score: 93.50 %
  - F2-score: 93.47 %

Per POS performance

- Tag: ADV Precision: 89.65 %, recall: 86.82 %, F1: 88.21
- Tag: DET Precision: 89.58 %, recall: 98.63 %, F1: 93.89
- Tag: VERB Precision: 95.25 %, recall: 90.59 %, F1: 92.86
- Tag: ADP Precision: 91.85 %, recall: 96.46 %, F1: 94.10
- Tag: CONJ Precision: 99.29 %, recall: 99.31 %, F1: 99.30
- Tag: PRON Precision: 92.90 %, recall: 97.81 %, F1: 95.29
- Tag: NUM Precision: 96.93 %, recall: 86.37 %, F1: 91.34
- Tag: ADJ Precision: 87.41 %, recall: 86.89 %, F1: 87.15
- Tag: X Precision: 65.55 %, recall: 45.02 %, F1: 53.38
- Tag: PRT Precision: 89.34 %, recall: 89.49 %, F1: 89.42
- Tag: NOUN Precision: 94.53 %, recall: 90.12 %, F1: 92.27
- Tag: . Precision: 97.55 %, recall: 99.89 %, F1: 98.70

Confusion Matrix (12 X 12)

Interpretation of confusion (error analysis)

- Noun-Det, Noun-Verb, Verb-Noun

- Same words are used in different senses in different sentences e.g. Play can be used as both noun and verb
- Noun and verbs have higher overall frequency in corpus

Data Processing Info (Pre-processing)

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

Inferencing/Decoding Info

- For each word, stored the maximum probability path ending at each tag from the list of Universal tags
- From tags for the last word of a sentence, selected tag with maximum probability since it has the maximum probability path for whole sentence and then used back pointer to find the path.

Any thoughts on generative vs. discriminative POS tagging

- Discriminative models model conditional probability P(Y|X) while generative models model joint probability P(Y, X)
- Hence, discriminative models don’t need to model marginal probability P(X)