Assignment 1
POS Tagger

Anup Kulkarni, Saurabh Sohoney, Prashanth Kamle

March 9, 2009
Implemented POS-Taggers

- Baseline Tagger – most frequent tag
- Bigram Tagger
  - Simple bigram tagger
  - Tagger with stemming
  - Tagger with beam search
- Trigram tagger
  - Simple bigram tagger
  - Tagger with stemming
  - Tagger with beam search
Strategy

- Training corpus: Brown-news category, simplified tags
- Viterbi and its Beam variation
- Testing using Four-fold cross-validation
- Morphology analysis: using Porter Stemmer
- Python programming language
Results

- Baseline: 81.65%
- Bigram: 89.48%
- Bigram, Stemmed: 86.95%
- Trigram: 85.76%
- Trigram, Stemmed: 79.98%
Per POS tag accuracy

- Baseline
- Bigram
- Stemmed bigram
- Trigram
- Stemmed trigram
Challenges faced

- **Unknown word handling**
  - Most likely tag corresponding to 3 character suffix of unknown word

- **Handling super small numbers (Underflow)**
  - Convert to log space
  - e.g. $\log(1.0\text{e}-205) = -205$

- **Handling log(0)**

- **Handling unknown tag sequences**
  - Assign equal probabilities to all possible tags
Beam search

- Restrict number of possible tags (states) at each level
- Keep a threshold probability log(0.2)
- Dynamic adjustment of Beam width
  - new_beam = old_beam + log(prev_max_prob)
References

- HMM-based Language-independent POS Tagger by Pradeep Varma D., Rakesh M., Ratna Sanyal, Indian Institute of Information Technology, Allahabad