You wanna predict tags? WE wanna predict tags, too!

Naoki Orii, Peter Schulam, & Ryan Carlson

The Task

- Get a bunch of tagged POS sequences

- Predict new ones

Our Approach

Ngrams

How Many Ngrams?!

Unigrams!

How Many Ngrams?!

Bigrams!!

How Many Ngrams?!

Trigrams!!!

How Many Ngrams?!

Fourgrams!!!!

How Many Ngrams?!

Fivegrams!!!!!

How Many Ngrams?!

Sixgrams!!!!!!

How Many Ngrams?!

Sevengrams!!!!!!!

How Many Ngrams?!

EIGHTgrams!!!!!!!!

Our Approach

- Ngrams

  - with psuedo counts

  - tried Good-Turing, but they didn't seem to work as well

Ngram

- NLTK? don't think so! (rolled our own since NLTK was returning probabilities > 1)

- Maximum likelihood estimates with psuedo-counts to avoid nil probabilities
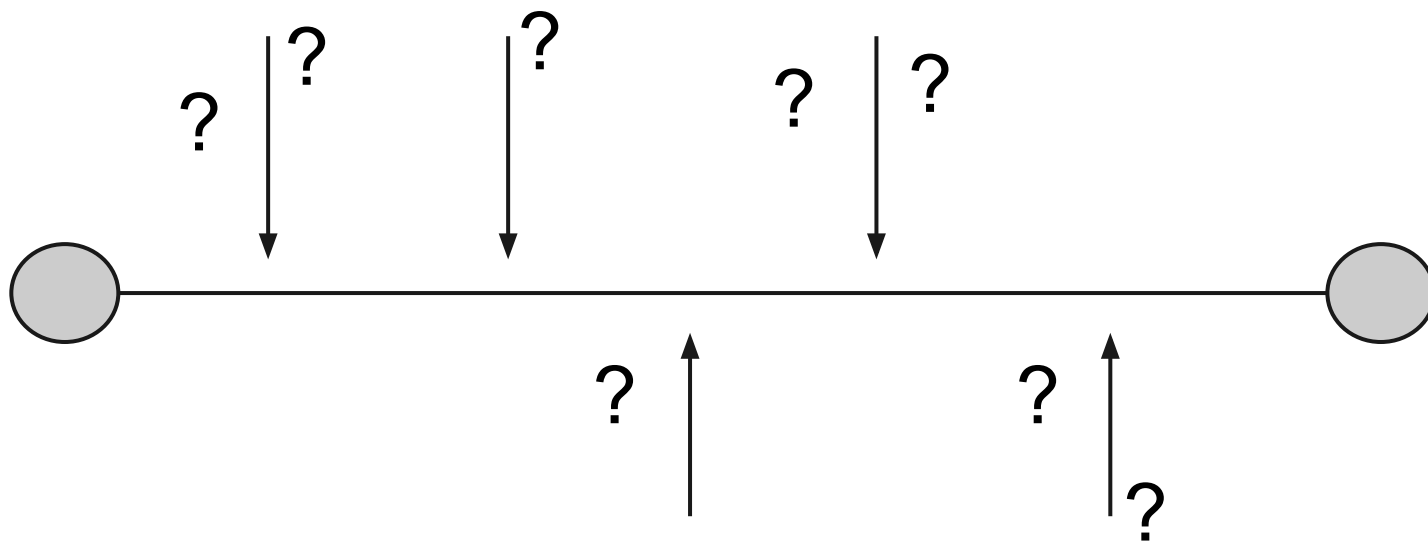- Interpolated on development set taken from chunk of training data

Our Approach (cont)

- Maxent
  - Lots of features
  - (explained soon)

max( )

Our Approach (cont)

- Linear Interpolation
  - optimize weights for each model wrt perplexity

Maxent

- Each created our own set of features

- **Peter**: shallow syntactic chunks, unigram, bigram, trigram

- **Naoki**: trigger pairs, distance-2 bigram and trigram

- **Ryan**: subset existence

Results

- Split into 90% train, 9% dev, 1% test
- Trained Models on *train* set
- Interpolated on *dev* set
- Evaluated on *test* set

- Accuracy: 37.59%, Perplexity: 8.107 bits/tag