CSCI 5832
Natural Language Processing

Lecture 15
Jim Martin

Today 3/8

- Review Sequence Labeling/Chunking
- More on Classifiers
- Break
- Project discussions

Statistical Sequence Labeling

- As with POS tagging, we can use rules to do partial parsing or we can train systems to do it for us. To do that we need training data and the right kind of encoding.
  - Training data
    - Hand tag a bunch of data (as with POS tagging)
    - Or even better, extract partial parse bracketing information from a treebank.

Encoding

- With the right encoding you can turn the labeled bracketing task into a tagging task. And then proceed exactly as we did with POS Tagging.
- We'll use whats called IOB labeling to do this.
  - I -> Inside
  - O -> Outside
  - B -> Begins

IOB encoding

The morning flight from Denver has arrived.
B_NP I_NP I_NP O B_NP O O

The morning flight from Denver has arrived
B_NP I_NP I_NP B_PP B_NP B_VP I_VP

- The first example shows the encoding for just base-NPs. There are 3 tags in this scheme.
- The second shows full coverage. In this scheme there are 2*N+1 tags. Where N is the number of constituents in your set.

Methods

- HMMs

$$\begin{aligned}\hat{T} &= \operatorname*{argmax}_T P(T|W) \\ &= \operatorname*{argmax}_T P(W|T)P(T) \\ &= \operatorname*{argmax}_T \prod_i P(word_i|tag_i) \prod_i P(tag_i|tag_{i-1})\end{aligned}$$

- Sequence Classification
  - Using any kind of standard ML-based classifier.