Sequence Labeling (II)

CS 690N, Spring 2018

Advanced Natural Language Processing

http://people.cs.umass.edu/~brenocon/anlp2018/

Brendan O’Connor

College of Information and Computer Sciences

University of Massachusetts Amherst

How to build a POS tagger?

- Sources of information:
  - POS tags of surrounding words: syntactic context
  - The word itself
  - Features, etc.!
    - Word-internal information
    - Features from surrounding words
    - External lexicons
    - Embeddings, LSTM states

← HMM

← Classifier

← CRF

Sequence labeling

- Seq. labeling as classification:
  Each position *m* gets an independent classification,
  as a log-linear model.

$$p(y_m \mid w_1..w_n)$$

$$\arg\max_y \theta^\mathsf{T} \mathbf{f}((\mathbf{w}, m), y)$$

$$\begin{aligned}
\boldsymbol{f}((\boldsymbol{w} = \textit{they can fish}, m = 1), \mathrm{N}) =& \langle \textit{they}, \mathrm{N} \rangle \\
\boldsymbol{f}((\boldsymbol{w} = \textit{they can fish}, m = 2), \mathrm{V}) =& \langle \textit{can}, \mathrm{V} \rangle \\
\boldsymbol{f}((\boldsymbol{w} = \textit{they can fish}, m = 3), \mathrm{V}) =& \langle \textit{fish}, \mathrm{V} \rangle.
\end{aligned}$$

Sequence labeling

- Seq. labeling as classification:
  Each position *m* gets an independent classification,
  as a log-linear model.

$$p(y_m \mid w_1..w_n)$$

$$\arg\max_y \theta^\mathsf{T} \mathbf{f}((\mathbf{w}, m), y)$$

$$\boldsymbol{f}((\boldsymbol{w} = \textit{they can fish}, m = 1), \mathrm{N}) = \langle \textit{they}, \mathrm{N} \rangle$$
$$\boldsymbol{f}((\boldsymbol{w} = \textit{they can fish}, m = 2), \mathrm{V}) = \langle \textit{can}, \mathrm{V} \rangle$$
$$\boldsymbol{f}((\boldsymbol{w} = \textit{they can fish}, m = 3), \mathrm{V}) = \langle \textit{fish}, \mathrm{V} \rangle.$$

- But syntactic (tag) context is sometimes necessary!

- Seq. labeling as log-linear ***structured prediction***

$$\hat{\boldsymbol{y}}_{1:M} = \underset{\boldsymbol{y}_{1:M} \in \mathcal{Y}(\boldsymbol{w}_{1:M})}{\operatorname{argmax}} \boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{w}_{1:M}, \boldsymbol{y}_{1:M}),$$

- Example: the ***Hidden Markov model***

$$p(\mathbf{w}, \mathbf{y}) = \prod_t p(y_t \mid y_{t-1}) p(w_t \mid y_t)$$

- Efficiently supports operations via dynamic programming –
because of **local (Markovian) assumptions**
  - P(w): Likelihood (generative model)
    - Forward algorithm
  - P(y | w): Predicted sequence ("decoding")
    - Viterbi algorithm
  - P(y_m | w): Predicted tag marginals
    - Forward-Backward algorithm
    - Supports EM for unsupervised HMM learning

Forward-Backward

- (handout)
- stopped 3/8 at the forward algorithm

Baum-Welch

- EM applied to HMMs
(where EM was really invented...)
- **E-step**: calculate marginals with forward-backward
  - $p(y_{t-1}, y_t \mid w_1..w_T)$
  - $p(y_t \mid w_1..w_T)$
- **M-step**: re-estimate parameters from expected counts
  - Transitions: will use pair marginals
  - Emissions: will use tag marginals

Viterbi algorithm

- If the feature function decomposes into local features, dynamic programming gives global solution

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, \boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{w}, \boldsymbol{y}) \qquad \boldsymbol{f}(\boldsymbol{w}, \boldsymbol{y}) = \sum_{m=1}^{M} \boldsymbol{f}(\boldsymbol{w}, y_m, y_{m-1}, m).$$

- Decompose:

$$\max_{\boldsymbol{y}} \boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{w}, \boldsymbol{y}) = \max_{\boldsymbol{y}_{1:M}} \sum_{m=1}^{M} \boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{w}, y_m, y_{m-1}, m)$$

- Define Viterbi variables:

$$v_m(k) \triangleq \max_{\boldsymbol{y}_{1:m-1}} \boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{w}, k, y_{m-1}, m) + \sum_{n=1}^{m-1} \boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{w}, y_n, y_{n-1}, n)$$

Viterbi algorithm

- If the feature function decomposes into local features, dynamic programming gives global solution

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, \boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{w}, \boldsymbol{y}) \qquad \boldsymbol{f}(\boldsymbol{w}, \boldsymbol{y}) = \sum_{m=1}^{M} \boldsymbol{f}(\boldsymbol{w}, y_m, y_{m-1}, m).$$

- Decompose:

$$\max_{\boldsymbol{y}} \boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{w}, \boldsymbol{y}) = \max_{\boldsymbol{y}_{1:M}} \sum_{m=1}^{M} \boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{w}, y_m, y_{m-1}, m)$$

$$= \max_{\boldsymbol{y}_{1:M}} \boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{w}, y_M, y_{M-1}, M) + \sum_{m=1}^{M-1} \boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{w}, y_m, y_{m-1}, m)$$

- Define Viterbi variables:

$$v_m(k) \triangleq \max_{\boldsymbol{y}_{1:m-1}} \boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{w}, k, y_{m-1}, m) + \sum_{n=1}^{m-1} \boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{w}, y_n, y_{n-1}, n)$$

Viterbi algorithm

- If the feature function decomposes into local features, dynamic programming gives global solution

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, \boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{w}, \boldsymbol{y}) \qquad \boldsymbol{f}(\boldsymbol{w}, \boldsymbol{y}) = \sum_{m=1}^{M} \boldsymbol{f}(\boldsymbol{w}, y_m, y_{m-1}, m).$$

- Decompose:

$$\begin{aligned}
\max_{\boldsymbol{y}} \boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{w}, \boldsymbol{y}) &= \max_{\boldsymbol{y}_{1:M}} \sum_{m=1}^{M} \boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{w}, y_m, y_{m-1}, m) \\
&= \max_{\boldsymbol{y}_{1:M}} \boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{w}, y_M, y_{M-1}, M) + \sum_{m=1}^{M-1} \boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{w}, y_m, y_{m-1}, m) \\
&= \max_{y_M} \max_{y_{M-1}} \boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{w}, y_M, y_{M-1}, M) + \max_{\boldsymbol{y}_{1:M-2}} \sum_{m=1}^{M-1} \boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{w}, y_m, y_{m-1}, m).
\end{aligned}$$

- Define Viterbi variables:

$$v_m(k) \triangleq \max_{\boldsymbol{y}_{1:m-1}} \boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{w}, k, y_{m-1}, m) + \sum_{n=1}^{m-1} \boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{w}, y_n, y_{n-1}, n)$$