TTIC 31190:
Natural Language Processing

Kevin Gimpel
Spring 2018

Lecture 11:
Part-of-Speech Tagging and other Sequence Labeling Tasks
Assignment 2 due today

• questions?
Project Proposal

• project proposal due May 9
• details have been posted
• groups of 2-3 are ok (but think about how you will divide up the work, especially with 3)
Project Details

• ideas:
  – replicate (part of) a published paper
  – apply NLP methods to a dataset or task related to your research
  – define a new NLP task/dataset

• if you’re working on a standard task, you do not need to have state of the art results

• but your project should be done carefully so that you can have confidence in your claims

• try to avoid a project that’s too ambitious
Project Report

- final report due June 6
  - May 30 for graduating students
- details forthcoming on project report format
Midterm

• midterm on Wednesday, May 16\textsuperscript{th}
• don’t worry about memorizing stuff
• we’ll give you most of the formulas/definitions you will need
Roadmap

• words, morphology, lexical semantics
• text classification
• language modeling
• word embeddings
• recurrent/recursive/convolutional networks in NLP
• sequence labeling, HMMs, dynamic programming
• syntax and syntactic parsing
• semantics, compositionality, semantic parsing
• machine translation and other NLP tasks
Encoders

- encoder: a function to represent a word sequence as a vector
- simplest: average word embeddings:

\[ f_{\text{avg}}(x) = \frac{1}{n} \sum_{i=1}^{n} \text{emb}(x_i) \]

- other choices: LSTMs, GRUs, CNNs, attention-weighted sum, etc.
Recurrent Neural Networks

Input is a sequence:

\[ \mathbf{x}_{t-1} \rightarrow \mathbf{h}_{t-1} \rightarrow \mathbf{h}_t \rightarrow \mathbf{h}_{t+1} \]

\[ \mathbf{x}_t \rightarrow \mathbf{h}_t \rightarrow \mathbf{h}_{t+1} \]

"hidden vector"
• so far, we’ve used RNNs to encode sequences for tasks like sequence classification
  – also used in translation, question answering, summarization, etc.
• but RNNs are also frequently used for generating sequences
“Output” Recurrent Neural Networks

\[ h_t = \tanh \left( W^{(x)} x_t + W^{(h)} h_{t-1} + b \right) \]

\[ y_t = \arg\max_{y \in \mathcal{O}} h_t^\top \text{emb}(y) \]
• \( y \) is a symbol, not a vector
• \( O \) is the “output” vocabulary
• we have a new parameter vector \( \text{emb}(y) \) for each output symbol in \( O \)
• \( \text{emb}(y) = x? \)
• probability distribution over output symbols?

\[
y_t = \arg\max_{y \in O} h_t^\top \text{emb}(y)
\]
\[ y_t = \arg\max_{y \in \mathcal{O}} h_t^\top \text{emb}(y) \]

\[ P(Y_t) = \text{softmax}(W h_t) \]

\[ W = \left[ \text{emb}(y_1)^\top ; \text{emb}(y_2)^\top ; \ldots ; \text{emb}(y_{|\mathcal{O}|})^\top \right] \]
Example: Language Modeling

... if the car ...

\[ x_{t-1} \rightarrow h_{t-1} \rightarrow y_{t-1} \]
\[ x_t \rightarrow h_t \rightarrow y_t \]
\[ x_{t+1} \rightarrow h_{t+1} \rightarrow y_{t+1} \]

- input: a word sequence
- output?
Example: Language Modeling

- target output at each position: next word in the sequence!
Language Modeling: Training

\[ -\log P(Y_{t-1} = ?) \]
Language Modeling: Training

... if

if

... the

the

car

car

$X_{t-1}$

$X_t$

$X_{t+1}$

$h_{t-1}$

$h_t$

$h_{t+1}$

$Y_{t-1}$

$Y_t$

$Y_{t+1}$

$- \log P(Y_{t-1} = \text{"the"}) - \log P(Y_t = \text{"car"})$ ...
• while we showed this for simple RNNs, it’s easy to instead use LSTMs, GRUs, etc.
• LSTMs/GRUs still produce a hidden vector at each position in the sequence, just like RNNs
• LSTM = most common choice for language modeling
Linguistic phenomena: summary so far...

• words have structure (stems and affixes)
• words have multiple meanings (senses) → word sense ambiguity
  – senses of a word can be homonymous or polysemous
  – senses have relationships:
    • synonymy, hyponymy (“is a”), meronymy (“part of”, “member of”)
• variability/flexibility of linguistic expression
  – many ways to express the same meaning (as you saw in Assignment 2)
  – word embeddings tell us when two words are similar
• today: part-of-speech
Part-of-Speech Tagging

Some questioned if Tim Cook’s first product would be a breakaway hit for Apple.
Some questioned if Tim Cook’s first product would be a breakaway hit for Apple.
Part-of-Speech (POS)

• functional category of a word:
  – noun, verb, adjective, etc.
  – how is the word functioning in its context?

• dependent on context like word sense, but different from sense:
  – sense represents word meaning, POS represents word function
  – sense uses a distinct category of senses per word, POS uses same set of categories for all words
<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
<th>Example</th>
<th>Tag</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>coordin. conjunction</td>
<td>and, but, or</td>
<td>SYM</td>
<td>symbol</td>
<td>+, %, &amp;</td>
</tr>
<tr>
<td>CD</td>
<td>cardinal number</td>
<td>one, two</td>
<td>TO</td>
<td>“to”</td>
<td>to</td>
</tr>
<tr>
<td>DT</td>
<td>determiner</td>
<td>a, the</td>
<td>UH</td>
<td>interjection</td>
<td>ah, oops</td>
</tr>
<tr>
<td>EX</td>
<td>existential ‘there’</td>
<td>there</td>
<td>VB</td>
<td>verb base form</td>
<td>eat</td>
</tr>
<tr>
<td>FW</td>
<td>foreign word</td>
<td>mea culpa</td>
<td>VBD</td>
<td>verb past form</td>
<td>ate</td>
</tr>
<tr>
<td>IN</td>
<td>preposition/sub-conj</td>
<td>of, in, by</td>
<td>VBG</td>
<td>verb gerund</td>
<td>eating</td>
</tr>
<tr>
<td>JJ</td>
<td>adjective</td>
<td>yellow</td>
<td>VBN</td>
<td>verb past participle</td>
<td>eaten</td>
</tr>
<tr>
<td>JJR</td>
<td>adj., comparative</td>
<td>bigger</td>
<td>VBP</td>
<td>verb non-3sg pres</td>
<td>eat</td>
</tr>
<tr>
<td>JJS</td>
<td>adj., superlative</td>
<td>wildest</td>
<td>VBZ</td>
<td>verb 3sg pres</td>
<td>eats</td>
</tr>
<tr>
<td>LS</td>
<td>list item marker</td>
<td>1, 2, One</td>
<td>WDT</td>
<td>wh-determiner</td>
<td>which, that</td>
</tr>
<tr>
<td>MD</td>
<td>modal</td>
<td>can, should</td>
<td>WP</td>
<td>wh-pronoun</td>
<td>what, who</td>
</tr>
<tr>
<td>NN</td>
<td>noun, sing. or mass</td>
<td>llama</td>
<td>WP$</td>
<td>possessive wh-</td>
<td>whose</td>
</tr>
<tr>
<td>NNS</td>
<td>noun, plural</td>
<td>llamas</td>
<td>WRB</td>
<td>wh-adverb</td>
<td>how, where</td>
</tr>
<tr>
<td>NNP</td>
<td>proper noun, sing.</td>
<td>IBM</td>
<td>$</td>
<td>dollar sign</td>
<td>$</td>
</tr>
<tr>
<td>NNPS</td>
<td>proper noun, plural</td>
<td>Carolinas</td>
<td>#</td>
<td>pound sign</td>
<td>#</td>
</tr>
<tr>
<td>PDT</td>
<td>predeterminer</td>
<td>all, both</td>
<td>“</td>
<td>left quote</td>
<td>‘ or “</td>
</tr>
<tr>
<td>POS</td>
<td>possessive ending</td>
<td>’s</td>
<td>”</td>
<td>right quote</td>
<td>’ or ”</td>
</tr>
<tr>
<td>PRP</td>
<td>personal pronoun</td>
<td>I, you, he</td>
<td>(</td>
<td>left parenthesis</td>
<td>[, (, {, &lt;</td>
</tr>
<tr>
<td>PRP$</td>
<td>possessive pronoun</td>
<td>your, one’s</td>
<td>)</td>
<td>right parenthesis</td>
<td>], ), }, &gt;</td>
</tr>
<tr>
<td>RB</td>
<td>adverb</td>
<td>quickly, never</td>
<td>,</td>
<td>comma</td>
<td>,</td>
</tr>
<tr>
<td>RBR</td>
<td>adverb, comparative</td>
<td>faster</td>
<td>.</td>
<td>sentence-final punc</td>
<td>. ! ?</td>
</tr>
<tr>
<td>RBS</td>
<td>adverb, superlative</td>
<td>fastest</td>
<td>:</td>
<td>mid-sentence punc</td>
<td>: ; ... − −</td>
</tr>
<tr>
<td>RP</td>
<td>particle</td>
<td>up, off</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Penn Treebank tag set
POS Ambiguity in Penn Treebank

• word that can be both noun and verb?
  – more often noun than verb:
    • increase: 248 NN vs. 127 VB (and 4 VBP)
    • place: 134 NN vs. 14 VB (and 4 VBP)

  – more often verb than noun:
    • makes: 182 VBZ vs. 5 NNS
    • transfer: 22 VB vs. 16 NN
POS Ambiguity in Penn Treebank

• word that can be both a singular noun and a plural noun?
  – “savings”, e.g.:

```
DT    NN    VBD    VBN    RB
The savings was given incorrectly ...

DT    JJ    NN    NN
a    Belgian savings bank
```
POS Ambiguity in Penn Treebank

• word that can be both a common noun and a proper noun?
  – “Earth”: 16 NNP vs. 5 NN
  – annotation inconsistencies: nothing in the context indicates which tag is used
  – these kinds of inconsistencies are common in annotated datasets, so it’s usually not possible to get perfect accuracy
POS Ambiguity in Penn Treebank

• word that can be both a common noun and a proper noun?
  – “Chapter”: 21 NNP vs. 41 NN
  – annotation inconsistencies:

    VB    VBG    IN    NNP    NNP    NN    NN
    consider filing for Chapter 11 bankruptcy protection

    NNP    VBD    IN    NN    CD    NN    NN
    Continental filed for Chapter 11 bankruptcy protection
How many tags can a word have?

words in Penn Treebank with the most unique tags:

7 down
6 that
6 set
6 put
6 open
6 hurt
6 cut
6 bet
6 back
5 vs.
5 the
5 spread
5 split
5 say
How many tags can a word have?

tag counts for down:

353 down RB
214 down RP
142 down IN
10 down JJ
1 down VBP
1 down RBR
1 down NN
How many tags can a word have?

tag counts for down:

353 down RB adverb
214 down RP particle
142 down IN preposition
10 down JJ adjective
1 down VBP verb (past tense)
1 down RBR comparative adverb
1 down NN singular noun
RP tag: particle

• test for verb particle:
• can you insert a modifier between the verb and its particle without it sounding weird?
  — take the trash out immediately
  — *take the trash immediately out
  — take the trash outside immediately
  — take the trash immediately outside

• out is a particle here, while outside is not
What about vs.? 

tag counts for vs.: 

15 vs. FW 
9 vs. IN 
6 vs. CC 
2 vs. NN 
1 vs. JJ
Universal Tag Set

• many use smaller sets of coarser tags
• e.g., “universal tag set” containing 12 tags:
  – noun, verb, adjective, adverb, pronoun, determiner/article, adposition (preposition or postposition), numeral, conjunction, particle, punctuation, other

![Example English sentence with its language specific and corresponding universal POS tags.](Petrov, Das, McDonald (2011))
• we removed some fine-grained POS tags, then added Twitter-specific tags:
  hashtag
  @-mention
  URL / email address
  emoticon
  Twitter discourse marker
  other (multi-word abbreviations, symbols, garbage)
• in Penn Treebank (1M words), word with most tags had 7 tags

• in Twitter POS-annotated data (40k words), word with most tags has how many tags?
How many tags can a word have?

words in Twitter with the most unique tags:

7 over
5 up
5 out
5 one
5 off
5 a
5 @
4 to
4 there
4 that
4 right
4 outside
4 no
4 n
How many tags can a word have?

words in Twitter with the most unique tags:

<table>
<thead>
<tr>
<th></th>
<th>over</th>
<th>4</th>
<th>there</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td></td>
<td>4</td>
<td>that</td>
</tr>
<tr>
<td>5</td>
<td>up</td>
<td>4</td>
<td>right</td>
</tr>
<tr>
<td>5</td>
<td>out</td>
<td>4</td>
<td>outside</td>
</tr>
<tr>
<td>5</td>
<td>one</td>
<td>4</td>
<td>no</td>
</tr>
<tr>
<td>5</td>
<td>off</td>
<td>4</td>
<td>n</td>
</tr>
</tbody>
</table>

Twitter shows a wider variety of uses for common words
## word sense vs. part-of-speech

<table>
<thead>
<tr>
<th>semantic or syntactic?</th>
<th>word sense</th>
<th>part-of-speech</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>semantic:</td>
<td>syntactic:</td>
</tr>
<tr>
<td></td>
<td>indicates meaning of word in its context</td>
<td>indicates function of word in its context</td>
</tr>
<tr>
<td>number of categories</td>
<td></td>
<td></td>
</tr>
<tr>
<td>inter-annotator</td>
<td></td>
<td></td>
</tr>
<tr>
<td>agreement</td>
<td></td>
<td></td>
</tr>
<tr>
<td>independent or joint</td>
<td></td>
<td></td>
</tr>
<tr>
<td>classification of nearby words?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### word sense vs. part-of-speech

<table>
<thead>
<tr>
<th></th>
<th>word sense</th>
<th>part-of-speech</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>semantic or syntactic?</strong></td>
<td>semantic: indicates meaning of word in its context</td>
<td>syntactic: indicates function of word in its context</td>
</tr>
<tr>
<td><strong>number of categories</strong></td>
<td>$</td>
<td>V</td>
</tr>
<tr>
<td><strong>inter-annotator agreement</strong></td>
<td>low; some sense distinctions are highly subjective</td>
<td>high; relatively few POS tags and function is relatively shallow / surface-level</td>
</tr>
<tr>
<td><strong>independent or joint classification of nearby words?</strong></td>
<td>independent: can classify a single word based on context words; structured prediction is rarely used</td>
<td>joint: strong relationship between tags of nearby words; structured prediction often used</td>
</tr>
</tbody>
</table>
How might POS tags be useful?

- text classification
- machine translation
- question answering
- speech synthesis (pronounce “contract”)
- ...

Models for POS Tagging

• today we’ll discuss simple models that do not use structured prediction
• these are often called “local” models
• they predict a tag for each word in a sequence, (and can use the entire word sequence to make each prediction)
• but they do not use information about previous predictions to make later predictions
• by contrast, structured prediction:
  – predict structures
  – or: make multiple predictions jointly
Feed-Forward Neural Networks for Twitter POS Tagging

ikr  smh  he  asked  fir  yo  last  name  so  he  can  add  u  on  fb  lololol

• in Assignment 3, you’ll build a neural network classifier to predict a word’s POS tag based on its context
• e.g., predict tag of yo given context
• what should the input $x$ be to the neural network?
  – it has to be independent of the label
  – it has to be a **fixed-length** vector
Feed-Forward Neural Networks for Twitter POS Tagging

• e.g., predict tag of yo given context
• what should the input $\mathbf{x}$ be?

$$\mathbf{x} = [0.4 \ 0.1 \ \ldots \ 0.9]^T$$

word vector for yo
Feed-Forward Neural Networks for Twitter POS Tagging

- e.g., predict tag of \( yo \) given context
- what should the input \( \mathbf{x} \) be?

\[
\mathbf{x} = \begin{bmatrix}
-0.2 & 0.5 & \ldots & 0.8 & 0.4 & 0.1 & \ldots & 0.9
\end{bmatrix}^T
\]

- word vector for \( fir \)
- word vector for \( yo \)
Feed-Forward Neural Networks for Twitter POS Tagging

• when using word vectors as part of input, we can also treat them as more parameters to be learned!
• this is called “updating” or “fine-tuning” the vectors (since they are initialized using something like word2vec)

\[ \mathbf{x} = [-0.2 \ 0.5 \ \ldots \ 0.8 \ 0.4 \ 0.1 \ \ldots \ 0.9]^\top \]
Feed-Forward Neural Networks for Twitter POS Tagging

- let’s use the center word + two words to the right:
  \[
  x = \begin{bmatrix}
  0.4 & \ldots & 0.9 \\
  0.2 & \ldots & 0.7 \\
  0.3 & \ldots & 0.6
  \end{bmatrix}^T
  \]

  – vector for \( yo \)
  – vector for \( last \)
  – vector for \( name \)

- if \( name \) is to the right of \( yo \), then \( yo \) is probably a form of \( your \)
- but our \( x \) above uses separate dimensions for each position!
  – i.e., \( name \) is two words to the right
  – what if \( name \) is one word to the right?
Feed-Forward Networks for POS Tagging

- feed-forward networks are OK for tagging
- they tend to work best with very small contexts (e.g., 1 word to left & right)
- with larger windows, probably not enough data to learn a good model
- also, distant words not very informative for POS tagging
- can also use convolutional networks defined on a window centered on the target word
RNNs for Part-of-Speech Tagging

... if                      the                     car ...”

- Input: a word sequence

Diagram:

- $x_{t-1}$
- $x_t$
- $x_{t+1}$
- $h_{t-1}$
- $h_t$
- $h_{t+1}$
- $y_{t-1}$
- $y_t$
- $y_{t+1}$
RNNs for Part-of-Speech Tagging

... if ... the ... car ...

• target output at each position: POS tag for corresponding word
RNN Taggers

• RNN POS taggers are simple and effective
• most common is to use some sort of bidirectional RNN, like a BiLSTM or BiGRU
RNN Taggers

• note: RNN taggers are not structured predictors

• yes, a structure is being predicted, but predictions for neighboring words are independent!

• BiRNN taggers do compute input representations that depend on the sentence context

• but they do not make any predictions jointly; each prediction is independent of all others
Sequence Labeling

• roughly: for each item in an input sequence, predict a label

• many sequence labeling tasks in NLP and other areas
  – computational biology, speech processing, video processing, etc.

• related class of tasks: segmentation, possibly with labeling of segments
Some questioned if Tim Cook’s first product would be a breakaway hit for Apple.
• there are many downloadable part-of-speech taggers and named entity recognizers:
  – Stanford POS tagger, NER labeler
  – TurboTagger, TurboEntityRecognizer
  – Illinois Entity Tagger
  – CMU Twitter POS tagger
  – Alan Ritter’s Twitter POS/NER labeler
They rarely seem to express any sort of shock, no matter what happens.
Some questioned if Tim Cook’s first product would be a breakaway hit for Apple.

Potential tags:

- ORGANIZATION
- LOCATION
- PERSON