Foundations of Natural Language Processing  
Lecture 8  
Part-of-speech Tagging and HMMs

Alex Lascarides  
(based on slides by Alex Lascarides, Sharon Goldwater & Philipp Koehn)  
9 February 2018

What is part of speech tagging?

• Given a string:
  
  This is a simple sentence

• Identify parts of speech (syntactic categories):
  
  This/DET is/VB a/DET simple/ADJ sentence/NOUN

Why do we care about POS tagging?

• POS tagging is a first step towards syntactic analysis (which in turn, is often useful for semantic analysis).
  
  – Simpler models and often faster than full parsing, but sometimes enough to be useful.
  
  – For example, POS tags can be useful features in text classification (see previous lecture) or word sense disambiguation (see later in course).

• Illustrates the use of hidden Markov models (HMMs), which are also used for many other tagging (sequence labelling) tasks.

Examples of other tagging tasks

• Named entity recognition: e.g., label words as belonging to persons, organizations, locations, or none of the above:
  
  Barack/PER Obama/PER spoke/NON from/NON the/NON White/LOC House/LOC today/NON ./NON

• Information field segmentation: Given specific type of text (classified advert, bibliography entry), identify which words belong to which “fields” (price/size/location, author/title/year)
  
  3BR/SIZE flat/TYPE in/NON Bruntsfield/LOC ./NON near/LOC main/LOC roads/LOC ./NON Bright/FEAT ./NON well/FEAT maintained/FEAT
Sequence labelling: key features

In all of these tasks, deciding the correct label depends on

- the word to be labeled
  - NER: *Smith* is probably a person.
  - POS tagging: *chair* is probably a noun.

- the labels of surrounding words
  - NER: if following word is an organization (say *Corp.*), then this word is more likely to be organization too.
  - POS tagging: if preceding word is a modal verb (say *will*) then this word is more likely to be a verb.

HMM combines these sources of information probabilistically.

Parts of Speech: reminder

- **Open class words** (or content words)
  - nouns, verbs, adjectives, adverbs
  - mostly content-bearing: they refer to objects, actions, and features in the world
  - *open* class, since there is no limit to what these words are, new ones are added all the time (*email*, *website*).

- **Closed class words** (or function words)
  - pronouns, determiners, prepositions, connectives, ...
  - there is a limited number of these
  - mostly functional: to tie the concepts of a sentence together

How many parts of speech?

- Both linguistic and practical considerations

- Corpus annotators decide. Distinguish between
  - proper nouns (names) and common nouns?
  - singular and plural nouns?
  - past and present tense verbs?
  - auxiliary and main verbs?
  - etc

- Commonly used tagsets for English usually have 40-100 tags. For example, the Penn Treebank has 45 tags.

<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><em>and</em>, <em>but</em>, <em>or</em></td>
<td>SYM</td>
<td>symbol</td>
<td>+, %, &amp;</td>
</tr>
<tr>
<td>CD</td>
<td>cardinal number</td>
<td><em>one</em>, <em>two</em></td>
<td>TO</td>
<td>&quot;to&quot;</td>
<td>to</td>
</tr>
<tr>
<td>DT</td>
<td>determiner</td>
<td><em>a</em>, <em>the</em></td>
<td>UH</td>
<td>interjection</td>
<td><em>ah</em>, <em>oops</em></td>
</tr>
<tr>
<td>EX</td>
<td>existential ‘there’</td>
<td><em>there</em></td>
<td>VB</td>
<td>verb base form</td>
<td><em>eat</em></td>
</tr>
<tr>
<td>FW</td>
<td>foreign word</td>
<td><em>mea culpa</em></td>
<td>VBD</td>
<td>verb past tense</td>
<td><em>ate</em></td>
</tr>
<tr>
<td>IN</td>
<td>preposition/sub-conj</td>
<td><em>of</em>, <em>in</em>, <em>by</em></td>
<td>VBG</td>
<td>verb gerund</td>
<td><em>eating</em></td>
</tr>
<tr>
<td>JJ</td>
<td>adjective</td>
<td><em>yellow</em></td>
<td>VBN</td>
<td>verb past participle</td>
<td><em>eaten</em></td>
</tr>
<tr>
<td>JJR</td>
<td>adj., comparative</td>
<td><em>bigger</em></td>
<td>VBP</td>
<td>verb non-3sg pres</td>
<td><em>eat</em></td>
</tr>
<tr>
<td>JJS</td>
<td>adj., superlative</td>
<td><em>wildest</em></td>
<td>VBZ</td>
<td>verb 3sg pres</td>
<td><em>eats</em></td>
</tr>
<tr>
<td>LS</td>
<td>list item marker</td>
<td><em>1</em>, <em>2</em>, <em>One</em></td>
<td>WDT</td>
<td>wh-determiner</td>
<td><em>which</em>, <em>that</em></td>
</tr>
<tr>
<td>MD</td>
<td>modal</td>
<td><em>can</em>, <em>should</em></td>
<td>WP</td>
<td>wh-pronoun</td>
<td><em>what</em>, <em>who</em></td>
</tr>
<tr>
<td>NN</td>
<td>noun, sing. or mass</td>
<td><em>llamas</em></td>
<td>WPS</td>
<td>possessive wh-</td>
<td><em>whose</em></td>
</tr>
<tr>
<td>NNS</td>
<td>noun, plural</td>
<td><em>llamas</em></td>
<td>WRB</td>
<td>wh-adverb</td>
<td><em>how</em>, <em>where</em></td>
</tr>
<tr>
<td>NNP</td>
<td>proper noun, sing.</td>
<td><em>IBM</em></td>
<td>$</td>
<td>dollar sign</td>
<td>$</td>
</tr>
<tr>
<td>NNPS</td>
<td>proper noun, plural</td>
<td><em>Carolinas</em></td>
<td>#</td>
<td>pound sign</td>
<td>#</td>
</tr>
<tr>
<td>PDT</td>
<td>preposition</td>
<td><em>all</em>, <em>both</em></td>
<td>‘s</td>
<td>right quote</td>
<td>’ or “</td>
</tr>
<tr>
<td>POS</td>
<td>possessive ending</td>
<td>‘s</td>
<td>“</td>
<td>left quote</td>
<td>’ or “</td>
</tr>
<tr>
<td>PRP</td>
<td>personal pronoun</td>
<td><em>I</em>, <em>you</em>, <em>he</em></td>
<td>(</td>
<td>left parenthesis</td>
<td>[, *, {, &lt;</td>
</tr>
<tr>
<td>PRPS</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><em>quickly</em>, <em>never</em></td>
<td>,</td>
<td>comma</td>
<td>,</td>
</tr>
<tr>
<td>RBR</td>
<td>adverb, comparative</td>
<td><em>faster</em></td>
<td>:</td>
<td>sentence-final punct</td>
<td>:</td>
</tr>
<tr>
<td>RBS</td>
<td>adverb, superlative</td>
<td><em>fastest</em></td>
<td>;</td>
<td>mid-sentence punct</td>
<td>;</td>
</tr>
<tr>
<td>RP</td>
<td>particle</td>
<td><em>up</em>, <em>off</em></td>
<td>:</td>
<td>continue</td>
<td>:</td>
</tr>
</tbody>
</table>

J&M Fig 5.6: Penn Treebank POS tags
POS tags in other languages

- Morphologically rich languages often have compound morphosyntactic tags
  \[ \text{Noun} + A3\text{sg} + P2\text{sg} + \text{Nom} \]  
  (J&M3, Chapter 10.7)
- Hundreds or thousands of possible combinations
- Predicting these requires more complex methods than what we will discuss (e.g., may combine an FST with a probabilistic disambiguation system)

Why is POS tagging hard?

The usual reasons!

- Ambiguity:
  - glass of water/NOUN vs. water/VERB
  - the plants lie/VERB down vs. tell a lie/NOUN
  - wind/VERB down vs. a mighty wind/NOUN (homographs)
  How about time flies like an arrow?
- Sparse data:
  - Words we haven’t seen before (at all, or in this context)
  - Word-Tag pairs we haven’t seen before (e.g., if we verb a noun)

Relevant knowledge for POS tagging

Remember, we want a model that decides tags based on

- The word itself
  - Some words may only be nouns, e.g. arrow
  - Some words are ambiguous, e.g. like, flies
  - Probabilities may help, if one tag is more likely than another
- Tags of surrounding words
  - two determiners rarely follow each other
  - two base form verbs rarely follow each other
  - determiner is almost always followed by adjective or noun

A probabilistic model for tagging

To incorporate these sources of information, we imagine that the sentences we observe were generated probabilistically as follows.

- To generate sentence of length \( n \):
  Let \( t_0 = <s> \)
  For \( i = 1 \) to \( n \)
    Choose a tag conditioned on previous tag: \( P(t_i|t_{i-1}) \)
    Choose a word conditioned on its tag: \( P(w_i|t_i) \)
- So, the model assumes:
  - Each tag depends only on previous tag: a bigram tag model.
  - Words are independent given tags
Probabilistic finite-state machine

• One way to view the model: sentences are generated by walking through states in a graph. Each state represents a tag.

Example transition probabilities

| $t_i$ | NNP | MD | VB | JJ | NN |...
|-------|-----|----|----|----|----|-------|
| $<s>$ | 0.2767 | 0.0006 | 0.0031 | 0.0453 | 0.0449 |...
| NNP   | 0.3777 | 0.0110 | 0.0009 | 0.0084 | 0.0584 |...
| MD    | 0.0008 | 0.0002 | 0.7968 | 0.0005 | 0.0008 |...
| VB    | 0.0322 | 0.0005 | 0.0050 | 0.0837 | 0.0615 |...
| JJ    | 0.0306 | 0.0004 | 0.0001 | 0.0733 | 0.4509 |...
| ...   | ...   | ... | ... | ... | ... |...

• Prob of moving from state $s$ to $s'$ (transition probability): $P(t_i = s' | t_{i-1} = s)$

Example transition probabilities

| $t_i$ | NNP | MD | VB | JJ | NN |...
|-------|-----|----|----|----|----|-------|
| $<s>$ | 0.2767 | 0.0006 | 0.0031 | 0.0453 | 0.0449 |...
| NNP   | 0.3777 | 0.0110 | 0.0009 | 0.0084 | 0.0584 |...
| MD    | 0.0008 | 0.0002 | 0.7968 | 0.0005 | 0.0008 |...
| VB    | 0.0322 | 0.0005 | 0.0050 | 0.0837 | 0.0615 |...
| JJ    | 0.0306 | 0.0004 | 0.0001 | 0.0733 | 0.4509 |...
| ...   | ...   | ... | ... | ... | ... |...

• Probabilities estimated from tagged WSJ corpus, showing, e.g.:
  - Proper nouns (NNP) often begin sentences: $P(\text{NNP}|<s>) \approx 0.28$
  - Modal verbs (MD) nearly always followed by bare verbs (VB).
  - Adjectives (JJ) are often followed by nouns (NN).

This table is incomplete!

• In the full table, every row must sum up to 1 because it is a distribution over the next state (given previous).

Probabilistic finite-state machine: outputs

• When passing through each state, emit a word.

| $t_i$ | NNP | MD | VB | JJ | NN |...
|-------|-----|----|----|----|----|-------|
| $<s>$ | 0.2767 | 0.0006 | 0.0031 | 0.0453 | 0.0449 |...
| NNP   | 0.3777 | 0.0110 | 0.0009 | 0.0084 | 0.0584 |...
| MD    | 0.0008 | 0.0002 | 0.7968 | 0.0005 | 0.0008 |...
| VB    | 0.0322 | 0.0005 | 0.0050 | 0.0837 | 0.0615 |...
| JJ    | 0.0306 | 0.0004 | 0.0001 | 0.0733 | 0.4509 |...
| ...   | ...   | ... | ... | ... | ... |...

• Prob of emitting $w$ from state $s$ (emission or output probability): $P(w_i = w | t_i = s)$
What can we do with this model?

• If we know the transition and output probabilities, we can compute the probability of a tagged sentence.

• That is,
  - suppose we have sentence $S = w_1 \ldots w_n$
  - and its tags $T = t_1 \ldots t_n$.
  - what is the probability that our probabilistic FSM would generate exactly that sequence of words and tags, if we stepped through at random?

This is the **joint probability** $P(S, T) = \prod_{i=1}^{n} P(t_i|t_{i-1})P(w_i|t_i)$
Example: computing joint prob. \( P(S, T) \)

What’s the probability of this tagged sentence?

This/DET is/VB a/DET simple/JJ sentence/NN

- First, add begin- and end-of-sentence \(<s>\) and \(</s>\). Then:

\[
p(S, T) = \prod_{i=1}^{n} P(t_i | t_{i-1}) P(w_i | t_i)
= P(\text{DET}<s>) P(\text{VB}|\text{DET}) P(\text{DET}|\text{VB}) P(JJ|\text{DET}) P(\text{NN}|JJ) P(</s>|\text{NN})
\cdot P(\text{This}|\text{DET}) P(\text{is}|\text{VB}) P(\text{a}|\text{DET}) P(\text{simple}|JJ) P(\text{sentence}|\text{NN})
\]

- Then, plug in the probabilities we estimated from our corpus.

Hidden Markov Model (HMM)

HMM is actually a very general model for sequences. Elements of an HMM:

- a set of states (here: the tags)
- a set of output symbols (here: words)
- initial state (here: beginning of sentence)
- state transition probabilities (here: \( p(t_i | t_{i-1}) \))
- symbol emission probabilities (here: \( p(w_i | t_i) \))

But... tagging?

Normally, we want to use the model to find the best tag sequence for an untagged sentence.

- Thus, the name of the model: hidden Markov model
  - Markov: because of Markov independence assumption (each tag/state only depends on fixed number of previous tags/states—here, just one).
  - hidden: because at test time we only see the words/emissions; the tags/states are hidden (or latent) variables.

- FSM view: given a sequence of words, what is the most probable state path that generated them?

Relationship to previous models

- N-gram model: a model for sequences that also makes a Markov assumption, but has no hidden variables.
- Naive Bayes: a model with hidden variables (the classes) but no sequential dependencies.
- HMM: a model for sequences with hidden variables.

Like many other models with hidden variables, we will use Bayes’ Rule to help us infer the values of those variables.

(We usually assume hidden variables are observed during training—annotated data; next class but we’ll discuss what happens if not.)
**Relationship to other models**

Side note for those interested:

- **Naive Bayes**: a generative model (use Bayes’ Rule, strong independence assumptions) for classification.

- **MaxEnt**: a discriminative model (model $P(y|x)$ directly, use arbitrary features) for classification.

- **HMM**: a generative model for sequences with hidden variables.

- **MEMM**: a discriminative model for sequences with hidden variables. Other sequence models can also use more features than HMM: e.g., Conditional Random Field; or see Toutanova et al. (2003) on web page.

**Formalizing the tagging problem**

Find the best tag sequence $T$ for an untagged sentence $S$:

$$\arg\max_T p(T|S)$$

- Bayes’ rule gives us:
  $$p(T|S) = \frac{p(S|T) p(T)}{p(S)}$$

- We can drop $p(S)$ if we are only interested in $\arg\max_T$:
  $$\arg\max_T p(T|S) = \arg\max_T p(S|T) p(T)$$

**Decomposing the model**

Now we need to compute $P(S|T)$ and $P(T)$ (actually, their product $P(S|T)P(T) = P(S, T)$).

- We already defined how!

- $P(T)$ is the state transition sequence:
  $$P(T) = \prod_i P(t_i|t_{i-1})$$

- $P(S|T)$ are the emission probabilities:
  $$P(S|T) = \prod_i P(w_i|t_i)$$
Search for the best tag sequence

- We have defined a model, but how do we use it?
  - given: word sequence \( S \)
  - wanted: best tag sequence \( T^* \)

- For any specific tag sequence \( T \), it is easy to compute \( P(S, T) = P(S|T)P(T) \).

\[
P(S|T) \ P(T) = \prod_i P(w_i|t_i) \ P(t_i|t_{i-1})
\]

- So, can’t we just enumerate all possible \( T \), compute their probabilities, and choose the best one?

Enumeration won’t work

- Suppose we have \( c \) possible tags for each of the \( n \) words in the sentence.

- How many possible tag sequences?

There are \( c^n \) possible tag sequences: the number grows exponentially in the length \( n \).

- For all but small \( n \), too many sequences to efficiently enumerate.

- This is starting to sound familiar...

The Viterbi algorithm

- As in min. edit distance, we’ll use a dynamic programming algorithm to solve the problem.

- The Viterbi algorithm finds the best tag sequence without explicitly enumerating all sequences.

- Like min. edit distance, the algorithm stores partial results in a chart to avoid recomputing them.

- Details next time.
The problem of finding the best tag sequence for a sentence is sometimes called **decoding**.

- Because, like spell correction etc, HMM can also be viewed as a noisy channel model.
  - Someone wants to send us a sequence of tags: \( P(T) \)
  - During encoding, “noise” converts each tag to a word: \( P(S|T) \)
  - We try to decode the observed words back to the original tags.

- In fact, **decoding** is a general term in NLP for inferring the hidden variables in a test instance (so, finding correct spelling of a misspelled word is also decoding).

**Summary**

- Part-of-speech tagging is a sequence labelling task.
- HMM uses two sources of information to help resolve ambiguity in a word’s POS tag:
  - The words itself
  - The tags assigned to surrounding words
- Can be viewed as a probabilistic FSM.
- Given a tagged sentence, easy to compute its probability. But finding the best tag sequence will need a clever algorithm.