

Part of Speech Tagging

COMP-599

Sept 29, 2015

Outline

Parts of speech in English

POS tagging as a sequence labelling problem

Markov chains revisited

Hidden Markov models

Parts of Speech in English

Nouns	<i>restaurant, me, dinner</i>
Verbs	<i>find, eat, is</i>
Adjectives	<i>good, vegetarian</i>
Prepositions	<i>in, of, up, above</i>
Adverbs	<i>quickly, well, very</i>
Determiners	<i>the, a, an</i>

What is a Part of Speech?

A kind of syntactic category that tells you some of the grammatical properties of a word.

The _____ was delicious.

- Only a noun fits here.

This hamburger is _____ than that one.

- Only a comparative adjective fits.

The cat ate. (OK – grammatical)

**The cat enjoyed.* (Ungrammatical. Note the *)

Important Note

You may have learned in grade school that nouns = things, verbs = actions. This is wrong!

Nouns that can be actions or events:

- *Examination, wedding, construction, opening*

Verbs that are not necessarily actions or events:

- *Be, have, want, enjoy, remember, realize*

Penn Treebank Tagset

CC	Coordinating conjunction	PRP\$	Possessive pronoun
CD	Cardinal number	RB	Adverb
DT	Determiner	RBR	Adverb, comparative
EX	Existential there	RBS	Adverb, superlative
FW	Foreign word	RP	Particle
IN	Preposition; subord. conjunct.	SYM	Symbol
JJ	Adjective	TO	to
JJR	Adjective, comparative	UH	Interjection
JJS	Adjective, superlative	VB	Verb, base form
LS	List item marker	VBD	Verb, past tense
MD	Modal	VBG	Verb, gerund or present part.
NN	Noun, singular or mass	VBN	Verb, past participle
NNS	Noun, plural	VBP	Verb, non-3rd pers. sing. pres.
NNP	Proper noun, singular	VBZ	Verb, 3rd pers. sing. pres.
NNPS	Proper noun, plural	WDT	Wh-determiner
PDT	Predeterminer	WP	Wh-pronoun
POS	Possessive ending	WP\$	Possessive wh-pronoun
PRP	Personal pronoun	WRB	Wh-adverb

Other Parts of Speech

Modals and auxiliary verbs

- *The police can and will catch the fugitives.*
- *Did the chicken cross the road?*

In English, these play an important role in question formation, and in specifying tense, aspect and mood.

Conjunctions

- *and, or, but, yet*

They connect and relate elements.

Particles

- *look up, turn down*

Can be parts of particle verbs. May have other functions (depending on what you consider a particle.)

Exercise

Give coarse POS tag labels to the following passage:

XPrize is a non-profit organization that designs public competitions to encourage technological development.

There are half a dozen XPrize competitions now underway, ranging from attempting a lunar landing to improving literacy in Africa.

Classifying Parts of Speech: Open Class

Open classes are parts of speech for which new words are readily added to the language (**neologisms**).

- Nouns *Twitter, Kleenex, turducken*
- Verbs *google, photoshop*
- Adjectives *Pastafarian, sick*
- Adverbs *automagically*
- Interjections *D'oh!*
- More at <http://neologisms.rice.edu/index.php>

Open class words usually convey most of the content. They tend to be **content words**.

Closed Class

Closed classes are parts of speech for which new words tend not to be added.

- Pronouns *I, he, she, them, their*
- Determiners *a, the*
- Quantifiers *some, all, every*
- Conjunctions *and, or, but*
- Modals and auxiliaries *might, should, ought*
- Prepositions *to, of, from*

Closed classes tend to convey grammatical information. They tend to be **function words**.

Corpus Differences

How fine-grained do you want your tags to be?

e.g., PTB tagset distinguishes singular from plural nouns

- NN *cat, water*
- NNS *cats*

e.g., PTB doesn't distinguish between **intransitive verbs** and **transitive verbs**

- VBD *listened* (intransitive)
- VBD *heard* (transitive)

Brown corpus (87 tags) vs. PTB (45)

Language Differences

Languages differ widely in which parts of speech they have, and in their specific functions and behaviours.

- In Japanese, there is no great distinction between nouns and pronouns. Pronouns are open class. OTTH, true verbs are a closed class.
 - I in Japanese: *watashi, watakushi, ore, boku, atashi, ...*
- In Wolof, verbs are not conjugated for person and tense. Instead, pronouns are.
 - *maa ngi* (1st person, singular, present continuous perfect)
 - *naa* (1st person, singular, past perfect)
- In Salishan languages (in the pacific northwest), there is no clear distinction between nouns and verbs.

POS Tagging

Assume we have a tagset and a corpus with words labelled with POS tags. What kind of problem is this?

Supervised or unsupervised?

Classification or regression?

Difference from classification that we saw last class—
context matters!

I saw the ...

The team won the match ...

Several cats ...

Sequence Labelling

Predict labels for *an entire sequence of inputs*:

? ? ? ? ? ? ? ? ?

Pierre Vinken , 61 years old , will join the board ...



NNP NNP , CD NNS JJ , MD VB DT NN

Pierre Vinken , 61 years old , will join the board ...

Must consider:

Current word

Previous context

Markov Chains

Our model will assume an underlying **Markov process** that generates the POS tags and words.

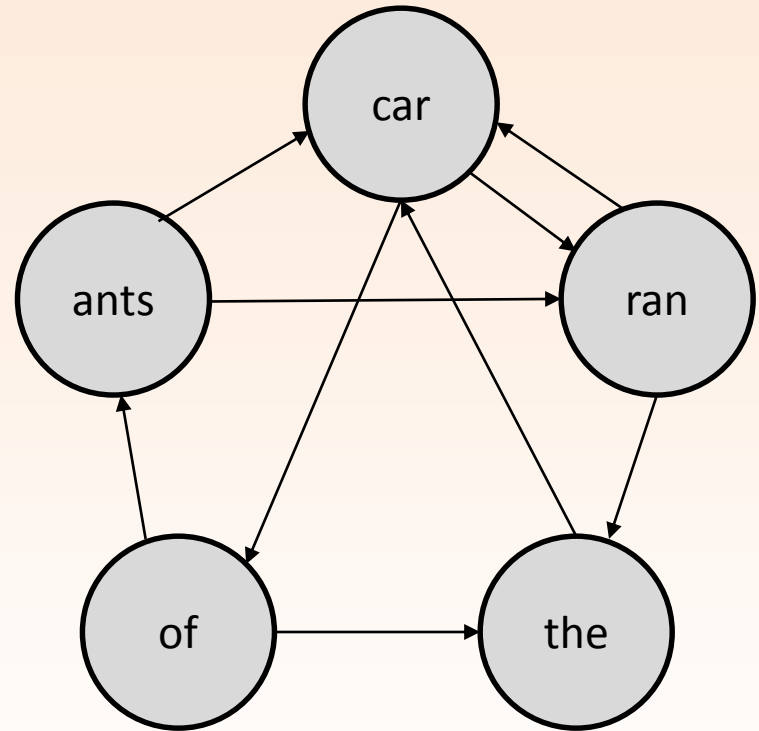
You've already seen Markov processes:

- Morphology: transitions between morphemes that make up a word
- N-gram models: transitions between words that make up a sentence

In other words, they are highly related to finite state automata

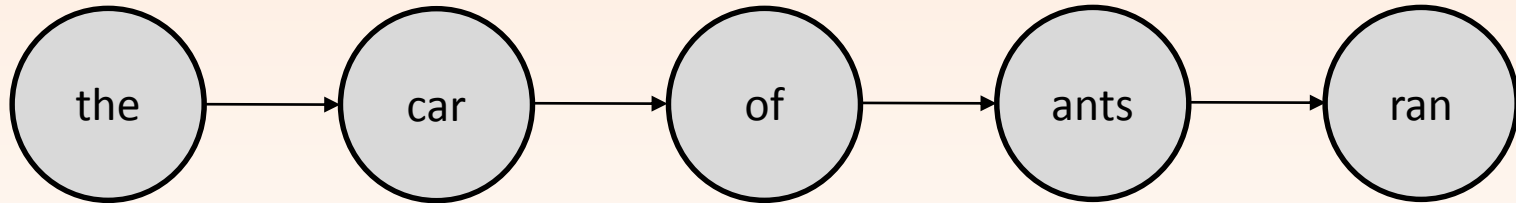
Observable Markov Model

- N states that represent unique observations about the world.
- Transitions between states are weighted—weights of all outgoing edges from a state sum to 1.
- e.g., this is a bigram model
- What would a trigram model look like?



Unrolling the Timesteps

A walk along the states in the Markov chain generates the text that is observed:



The probability of the observation is the product of all the edge weights (i.e., transition probabilities).

Hidden Variables

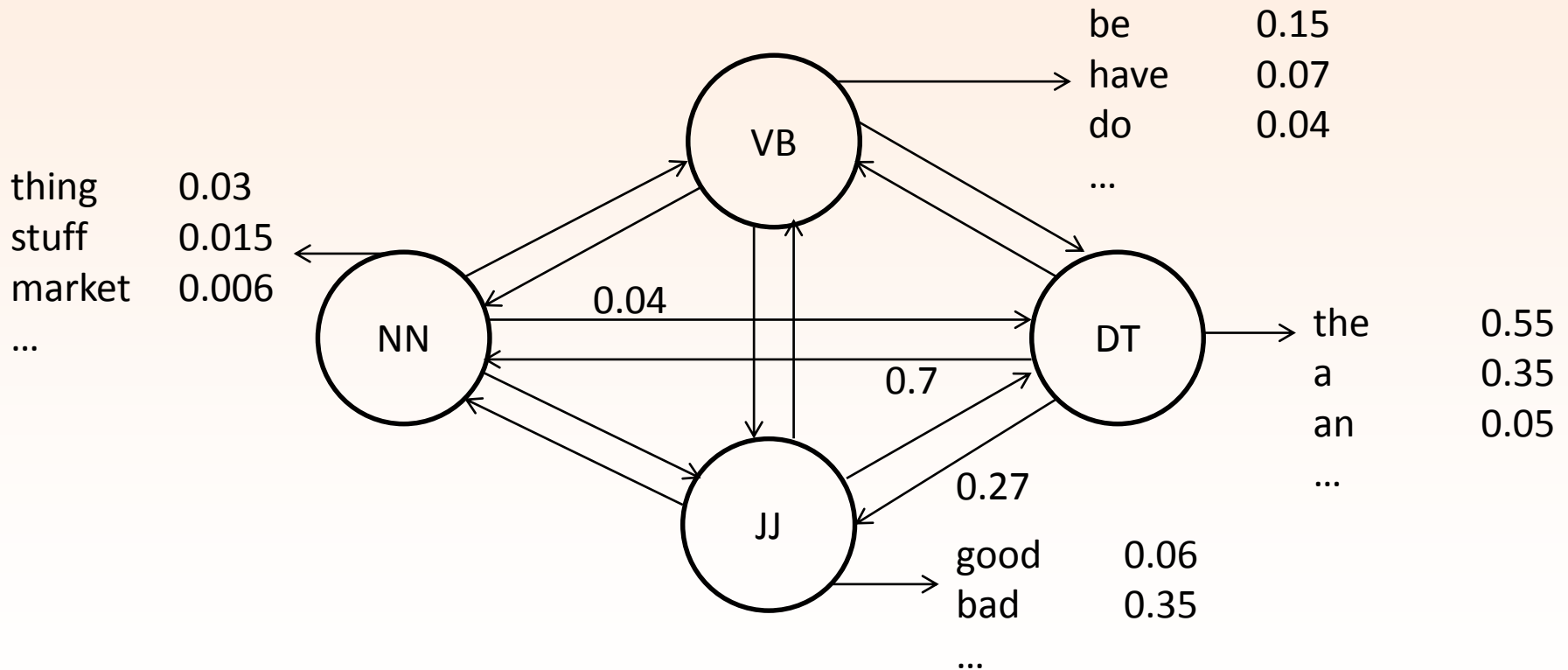
The POS tags to be predicted are **hidden variables**. We don't see them during test time (and sometimes not during training either).

It is very common to have hidden phenomena:

- Encrypted symbols are outputs of hidden messages
- Genes are outputs of functional relationships
- Weather is the output of hidden climate conditions
- Stock prices are the output of market conditions
- ...

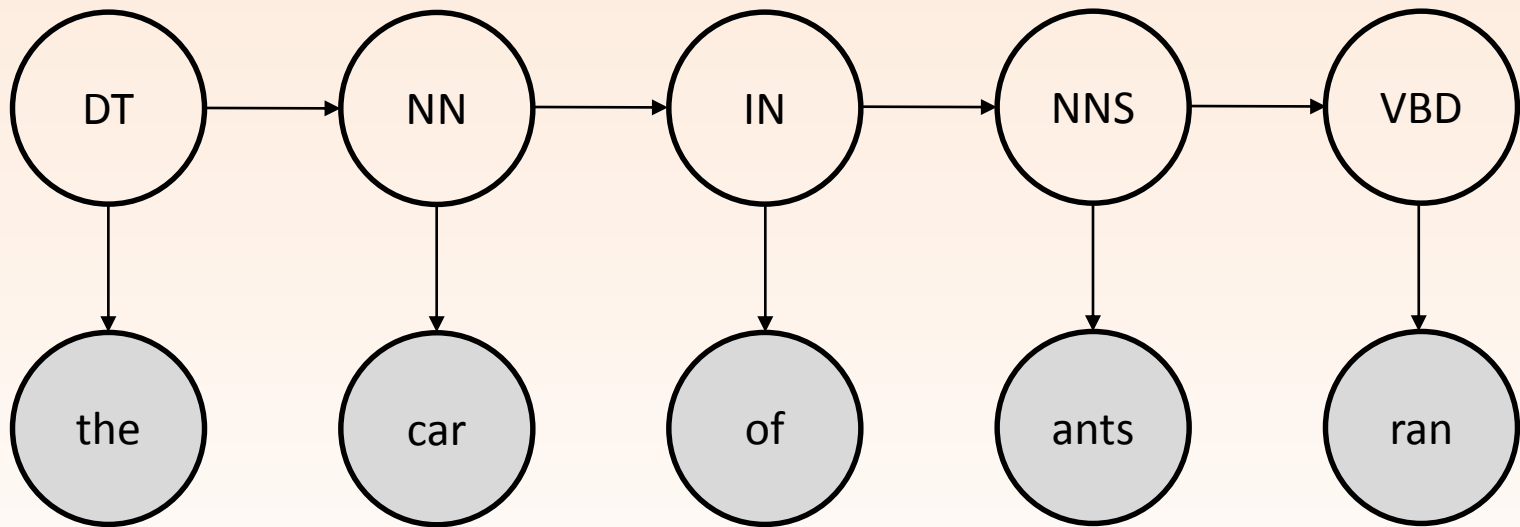
Markov Process w/ Hidden Variables

Model transitions between *POS tags*, and outputs (“emits”) a word which is observed at each timestep.



Unrolling the Timesteps

Now, the sample looks something like this:



Probability of a Sequence

Suppose we know *both* the sequence of POS tags and words generated by them:

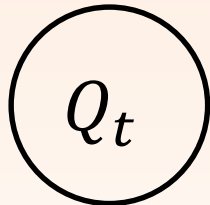
$$\begin{aligned} &P(\textit{The}/DT \textit{ car}/NN \textit{ of}/IN \textit{ ants}/NNS \textit{ ran}/VBD) \\ &= P(DT) \times P(DT \xrightarrow{\textit{emit}} \textit{Seize}) \\ &\quad \times P(DT \xrightarrow{\textit{trans}} NN) \times P(NN \xrightarrow{\textit{emit}} \textit{car}) \\ &\quad \times P(NN \xrightarrow{\textit{trans}} IN) \times P(IN \xrightarrow{\textit{emit}} \textit{of}) \\ &\quad \times P(IN \xrightarrow{\textit{trans}} NNS) \times P(NNS \xrightarrow{\textit{emit}} \textit{ants}) \\ &\quad \times P(NNS \xrightarrow{\textit{trans}} VBD) \times P(VBD \xrightarrow{\textit{emit}} \textit{ran}) \end{aligned}$$

- Product of hidden state *transitions* and observation *emissions*
- Note independence assumptions

Graphical Models

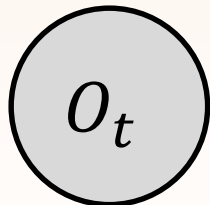
Since we now have many random variables, it helps to visualize them graphically. Graphical models precisely tell us:

- Latent or hidden random variables (clear)



$P(Q_t = VB)$: Probability that t^{th} tag is VB

- Observed random variables (filled)

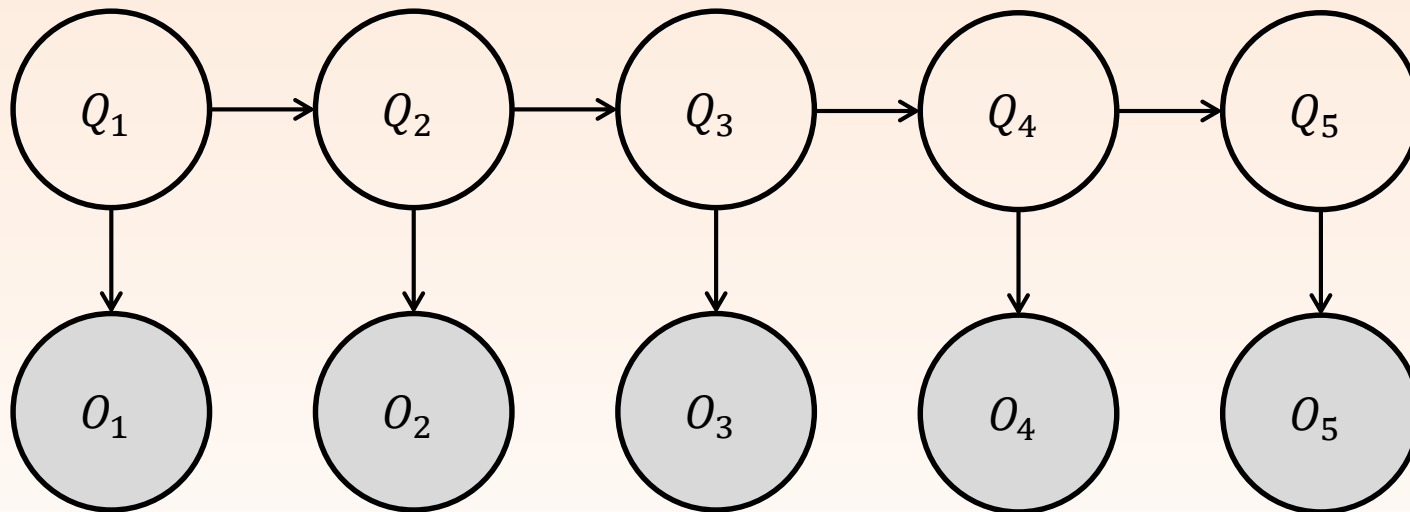


$P(O_t = \textit{ants})$: Probability that t^{th} word is *ants*

- Conditional independence assumptions (the edges)

Hidden Markov Models

Graphical representation

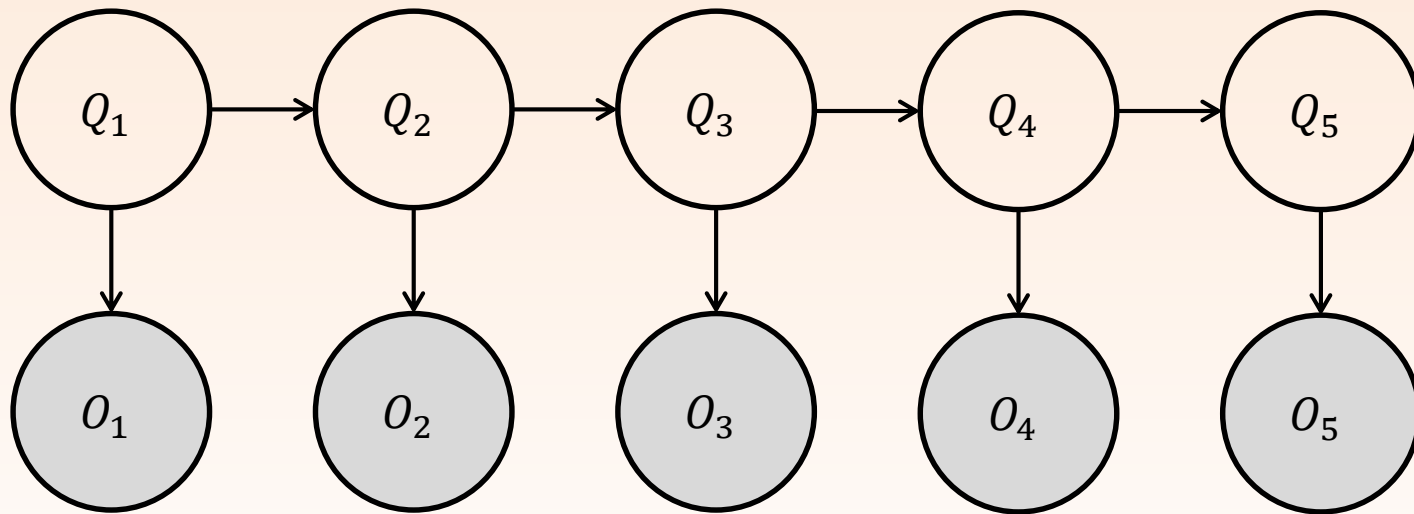


Denote entire sequence of tags as \mathbf{Q}

Entire sequence of words as \mathbf{O}

Decomposing the Joint Probability

Graph specifies how joint probability decomposes



$$P(\mathbf{O}, \mathbf{Q}) = P(Q_1) \prod_{t=1}^{T-1} P(Q_{t+1}|Q_t) \prod_{t=1}^T P(O_t|Q_t)$$

Initial state probability

State transition probabilities

Emission probabilities

Model Parameters

Let there be N possible tags, W possible words

Parameters θ has three components:

1. Initial probabilities for Q_1 :

$$\Pi = \{\pi_1, \pi_2, \dots, \pi_N\} \quad (\text{categorical})$$

2. Transition probabilities for Q_t to Q_{t+1} :

$$A = \{a_{ij}\} \quad i, j \in [1, N] \quad (\text{categorical})$$

3. Emission probabilities for Q_t to O_t :

$$B = \{b_i(w_k)\} \quad i \in [1, N], k \in [1, W] \quad (\text{categorical})$$

How many distributions and values of each type are there?

Model Parameters' MLE

Recall categorical distributions' MLE:

$$P(\text{outcome } i) = \frac{\#(\text{outcome } i)}{\#(\text{all events})}$$

For our parameters:

$$\pi_i = P(Q_1 = i) = \frac{\#(Q_1 = i)}{\#(\text{sentences})}$$

$$a_{ij} = P(Q_{t+1} = j | Q_t = i) = \#(i, j) / \#(i)$$

$$b_{ik} = P(O_t = k | Q_t = i) = \#(\text{word } k, \text{ tag } i) / \#(i)$$

Exercise in Supervised Training

What are the MLE for the following training corpus?

DT NN VBD IN DT NN

the cat sat on the mat

DT NN VBD JJ

the cat was sad

RB VBD DT NN

so was the mat

DT JJ NN VBD IN DT JJ NN

the sad cat was on the sad mat

Smoothing

Our previous discussion about smoothing and OOV items applies here too!

- Can smooth all the different types of distributions
- Recall this is called the MAP estimate

Next Time

- Now that we have a model, how do we actually tag a new sentence?
- What about unsupervised and semi-supervised learning?