Part-of-Speech (POS) tagging

Part-of-Speech (POS) tagging

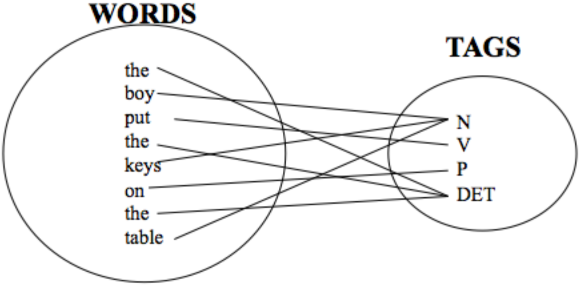
Task

Given a text of English, identify the parts of speech of each word

Part-of-Speech (POS) tagging

Task

Given a text of English, identify the parts of speech of each word



Parts of Speech: How many?

Open class words (content words)

- nouns, verbs, adjectives, adverbs
- mostly content-bearing: they refer to objects, actions, and features in the world
- *open class*, since new words are added all the time

E.g., new words like 'googling', 'photoshop', etc. get added to English

Parts of Speech: How many?

Open class words (content words)

- nouns, verbs, adjectives, adverbs
- mostly content-bearing: they refer to objects, actions, and features in the world
- *open class*, since new words are added all the time

Closed class words

- pronouns, determiners, prepositions, connectives, ...
- there is a limited number of these
- *mostly functional:* to tie the concepts of a sentence together

POS examples

N	noun	chair, bandwidth, pacing
V	verb	study, debate, munch
ADJ	adj	purple, tall, ridiculous
ADV	adverb	unfortunately, slowly,
P	preposition	of, by, to
PRO	pronoun	I, me, mine
DET	determiner	the, a, that, those

POS tagging: Choosing a tagset

- To do POS tagging, a standard set needs to be chosen

POS tagging: Choosing a tagset

- To do POS tagging, a standard set needs to be chosen
- Could pick very coarse tagsets
  N, V, Adj, Adv

POS tagging: Choosing a tagset

- To do POS tagging, a standard set needs to be chosen
- Could pick very coarse tagsets
  N, V, Adj, Adv
- More commonly used set is finer grained, "UPenn TreeBank tagset", 45 tags

UPenn TreeBank POS tag set

Variations of adjectives

Variations of nouns

Variations of verbs

Tag	Description	Example	Tag	Description	Example
CC	Coordin. Conjunction	*and, but, or*	SYM	Symbol	+,%, &
CD	Cardinal number	*one, two, three*	TO	"to"	*to*
DT	Determiner	*a, the*	UH	Interjection	*ah, oops*
EX	Existential 'there'	*there*	VB	Verb, base form	*eat*
FW	Foreign word	*mea culpa*	VBD	Verb, past tense	*ate*
IN	Preposition/sub-conj	*of, in, by*	VBG	Verb, gerund	*eating*
JJ	Adjective	*yellow*	VBN	Verb, past participle	*eaten*
JJR	Adj., comparative	*bigger*	VBP	Verb, non-3sg pres	*eat*
JJS	Adj., superlative	*wildest*	VBZ	Verb, 3sg pres	*eats*
LS	List item marker	*1, 2, One*	WDT	Wh-determiner	*which, that*
MD	Modal	*can, should*	WP	Wh-pronoun	*what, who*
NN	Noun, sing. or mass	*llama*	WP$	Possessive wh-	*whose*
NNS	Noun, plural	*llamas*	WRB	Wh-adverb	*how, where*
NNP	Proper noun, singular	*IBM*	$	Dollar sign	*$*
NNPS	Proper noun, plural	*Carolinas*	#	Pound sign	*#*
PDT	Predeterminer	*all, both*	"	Left quote	(' or ")
POS	Possessive ending	*'s*	"	Right quote	(' or ")
PRP	Personal pronoun	*I, you, he*	(	Left parenthesis	([, (, {, <)
PRP$	Possessive pronoun	*your, one's*	)	Right parenthesis	( ],), }, >)
RB	Adverb	*quickly, never*	,	Comma	,
RBR	Adverb, comparative	*faster*	.	Sentence-final punc	(. ! ?)
RBS	Adverb, superlative	*fastest*	:	Mid-sentence punc	(: ; ... – -)
RP	Particle	*up, off*			

Using the UPenn tagset

Example Sentence

The grand jury commented on a number of other topics.

Using the UPenn tagset

Example Sentence

The grand jury commented on a number of other topics.

POS tagged sentence

The/DT grand/JJ jury/NN commmented/VBD on/IN a/DT number/NN of/IN other/JJ topics/NNS ./.

Why is POS tagging hard?

Main reason: the same word can have different POS tags depending on the context in which it is used

Why is POS tagging hard?

Words often have more than one POS: back

- The back door:

Why is POS tagging hard?

Words often have more than one POS: back

- The back door: *back/JJ*
- On my back:

Why is POS tagging hard?

Words often have more than one POS: back

- The back door: *back/JJ*
- On my back: *back/NN*
- Win the voters back:

Why is POS tagging hard?

Words often have more than one POS: back

- The back door: *back/JJ*
- On my back: *back/NN*
- Win the voters back: *back/RB* **Adverb**
- Promised to back the bill:

Why is POS tagging hard?

Words often have more than one POS: back

- The back door: *back/JJ*
- On my back: *back/NN*
- Win the voters back: *back/RB*
- Promised to back the bill: *back/VB*

Why is POS tagging hard?

Words often have more than one POS: back

- The back door: *back/JJ*
- On my back: *back/NN*
- Win the voters back: *back/RB*
- Promised to back the bill: *back/VB*

POS tagging problem

To determine the POS tag for a particular instance of a word

How common is the problem of a word having ambiguous POS tags?

Ambiguous word types in the Brown Corpus

Ambiguity in the Brown corpus

- 40% of word tokens are ambiguous
- 12% of word types are ambiguous **12% distinct words**

Ambiguous word types in the Brown Corpus

Ambiguity in the Brown corpus

- 40% of word tokens are ambiguous
- 12% of word types are ambiguous
- Breakdown of ambiguous word types:

Unambiguous (1 tag)	35,340
Ambiguous (2–7 tags)	4,100
2 tags	3,760
3 tags	264
4 tags	61
5 tags	12
6 tags	2
7 tags	1 ("still")

Number of distinct words in the Brown corpus that have k tags

How bad is the ambiguity problem?

- One tag is usually more likely than the others.
  In the Brown corpus, *race* is a noun 98% of the time, and a verb 2% of the time

How bad is the ambiguity problem?

- One tag is usually more likely than the others.
  In the Brown corpus, *race* is a noun 98% of the time, and a verb 2% of the time
- A tagger for English that simply chooses the most likely tag for each word can achieve good performance

How bad is the ambiguity problem?

- One tag is usually more likely than the others.
  In the Brown corpus, *race* is a noun 98% of the time, and a verb 2% of the time
- A tagger for English that simply chooses the most likely tag for each word can achieve good performance
- Any new approach should be compared against the unigram baseline (assigning each token to its most likely tag)

Deciding the correct POS

Developing a gold standard for POS (for evaluating algorithms) can itself be challenging for some types of text / corpora.

Detailed manuals may be needed for the annotators.

Can be difficult even for people

- Mrs./NNP Shaefer/NNP never/RB got/VBD around/_ to/TO joining/VBG.
- All/DT we/PRP gotta/VBN do/VB is/VBZ go/VB around/_ the/DT corner/NN.
- Chateau/NNP Petrus/NNP costs/VBZ around/_ 2500/CD.

Relevant knowledge for POS tagging

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

Relevant knowledge for POS tagging

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

Local context

- Two determiners rarely follow each other
- Two base form verbs rarely follow each other
- Determiner is almost always followed by adjective or noun

Need two types of knowledge to understand POS tag of a word —
(i) knowledge about the word, e.g., most likely POS tag
(ii) knowledge of the context in which the word has been used

POS tagging: Two approaches

Rule-based Approach

- Assign each word in the input a list of potential POS tags
- Then winnow down this list to a single tag using hand-written rules

Knowledge-driven approach, e.g., using knowledge of English grammar.

POS tagging: Two approaches

Rule-based Approach

- Assign each word in the input a list of potential POS tags
- Then winnow down this list to a single tag using hand-written rules

Statistical tagging

- Get a training corpus of tagged text, learn the transformation rules from the most frequent tags (TBL tagger)
- Probabilistic: Find the most likely sequence of tags T for a sequence of words W

TBL - Transformation Based Learning; learn transformation rules over POS tags

TBL Tagger

Assume we have a training set where the words have been tagged with their correct (most likely) tags.

Label the training set with most frequent tags

- The can was rusted.
- The/DT can/MD was/VBD rusted/VBD.

This is our first guess (most frequent tag for each word).

The guess is correct for ‘The’ and ‘was’, but incorrect for ‘can’ and ‘rusted’. From the training corpus, we know that the correct tag for ‘can’ is NN and the correct tag for ‘rusted’ is VBN.

Add some transformation rules to correct these mistakes.

TBL Tagger

Label the training set with most frequent tags

- The can was rusted.
- The/DT can/MD was/VBD rusted/VBD.

Add transformation rules to reduce training mistakes

- MD →NN: DT_ **MD should be changed to NN, if preceded by DT**
- VBD→VBN: VBD_

We will frame such rules whenever our predicted tag does not match the gold standard tag (in the training data).

Probabilistic Tagging: Two different families of models

Problem at hand

We have some data $\{(d,c)\}$ of paired observations d and hidden classes c.

Probabilistic Tagging: Two different families of models

Problem at hand

We have some data $\{(d,c)\}$ of paired observations d and hidden classes c.

Different instances of d and c

- **Part-of-Speech Tagging**:

Probabilistic Tagging: Two different families of models

Problem at hand

We have some data $\{(d, c)\}$ of paired observations d and hidden classes c.

Different instances of d and c

- **Part-of-Speech Tagging**: words are observed and tags are hidden.
- **Text Classification**:

Probabilistic Tagging: Two different families of models

Problem at hand

We have some data $\{(d,c)\}$ of paired observations d and hidden classes c.

Different instances of d and c

- **Part-of-Speech Tagging**: words are observed and tags are hidden.
- **Text Classification**: sentences/documents are observed and the category is hidden.
  Categories can be positive/negative for sentiments ..
  sports/politics/business for documents ...

Probabilistic Tagging: Two different families of models

Problem at hand

We have some data $\{(d,c)\}$ of paired observations d and hidden classes c.

Different instances of d and c

- **Part-of-Speech Tagging**: words are observed and tags are hidden.
- **Text Classification**: sentences/documents are observed and the category is hidden.
  Categories can be positive/negative for sentiments ..
  sports/politics/business for documents ...

What gives rise to the two families?

Whether they generate the observed data from hidden stuff or the hidden structure given the data?

Generative vs. Conditional Models

Generative (Joint) Models

Generate the observed data from hidden stuff, i.e. put a probability over the observations given the class: $P(d,c)$ in terms of $P(d|c)$

Generative vs. Conditional Models

Generative (Joint) Models

Generate the observed data from hidden stuff, i.e. put a probability over the observations given the class: $P(d,c)$ in terms of $P(d|c)$
e.g. Naïve Bayes' classifiers, Hidden Markov Models etc.

Generative vs. Conditional Models

Generative (Joint) Models

Generate the observed data from hidden stuff, i.e. put a probability over the observations given the class: $P(d,c)$ in terms of $P(d|c)$
e.g. Naïve Bayes' classifiers, Hidden Markov Models etc.

Discriminative (Conditional) Models

Take the data as given, and put a probability over hidden structure given the data: $P(c|d)$

Generative vs. Conditional Models

Generative (Joint) Models

Generate the observed data from hidden stuff, i.e. put a probability over the observations given the class: $P(d,c)$ in terms of $P(d|c)$
e.g. Naïve Bayes' classifiers, Hidden Markov Models etc.

Discriminative (Conditional) Models

Take the data as given, and put a probability over hidden structure given the data: $P(c|d)$
e.g. Logistic regression, maximum entropy models, conditional random fields

Generative vs. Conditional Models

Generative (Joint) Models

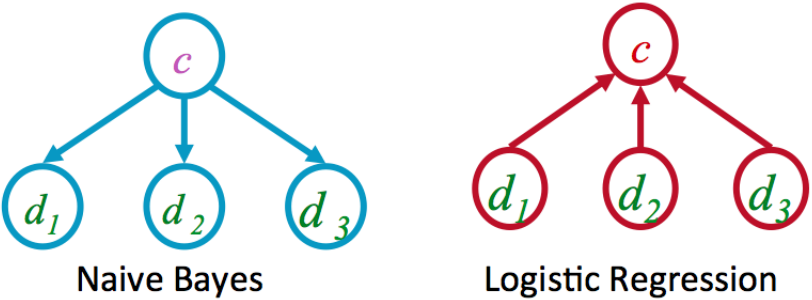
Generate the observed data from hidden stuff, i.e. put a probability over the observations given the class: $P(d,c)$ in terms of $P(d|c)$
e.g. Naïve Bayes' classifiers, Hidden Markov Models etc.

Discriminative (Conditional) Models

Take the data as given, and put a probability over hidden structure given the data: $P(c|d)$
e.g. Logistic regression, maximum entropy models, conditional random fields

SVMs, perceptron, etc. are discriminative classifiers but not directly probabilistic

Generative vs. Discriminative Models



Example problem: Classify documents into a set of classes, e.g., politics, religion, sports, entertainment, …

Discriminative approach: given a document, which is the most probable class?

Generative approach: given a document, from which class is this document most likely to have been generated?
E.g., suppose you have a Language Model for each class.