ANTONIS ANASTASOPOULOS
CS499 INTRODUCTION TO NLP

PART-OF-SPEECH TAGGING

https://cs.gmu.edu/~antonis/course/cs499-spring21/

With adapted slides by David Mortensen and Alan Black

STRUCTURE OF THIS LECTURE

1. The POS-tagging task
2. Parts of speech
3. POS-tagging with count-based models
4. Neural POS Tagging

POS TAGGING

POS TAGGING

My cat who lives dangerously no longer has nine lives.

POS TAGGING

My cat who **lives** dangerously no longer has nine **lives**.

lives: noun /lajvz/
lives: verb /livz/

The task is:

> Input: a sequence of word tokens $\mathbf{w}$
> Output: a sequence of part-of-speech tags $\mathbf{t}$ (one per word)

Note: the linguistic facts are considerably more complicated that the assumptions of the structure of this task, but there are good reasons for keeping it simple.

EXAMPLE

	Charlie	Brown	received	a	valentine	.
POS						
Features						

WHY HAVE PARTS OF SPEECH

There are too many words

- You'd need a lot of data to train rules
- Due to sparsity, rules would be very specific

PoS tags allow models to generalize

Give useful reduction in model sizes

There are many different tag sets

- You want the right one for your task

PARTS OF SPEECH

“

SO YOU WORK ON POS TAGGING. WHAT’S A PART OF SPEECH?

— David Kaplan

”

WHAT ARE PARTS OF SPEECH

The lexicon (collection of words of a language) is **not** some amorphous soup

It is somewhat soup-like, but it is a chunky soup:

- Small, finite number of categories
- Structured subcategories within these categories
- These categories are *soft*

If you ignore the structured nature of the lexicon, you are making life hard for yourself!

WHAT ARE PARTS OF SPEECH

A limited number of tags for word “class”

Distributional

- has the same contexts
- has the same syntactic functions (subj, obj, mod)
- occurs in same positions in syntactic structure

Morphological

- allows the same suffixes, prefixes

Not about meaning

- your English teacher lied to you — get used to it

OPEN-CLASS PARTS OF SPEECH

ENGLISH NOUNS

Can be subjects and objects of verbs

- *This book is about geography.*
- *I read a good book.*

Can be objects of prepositions

- *I'm mad about book*

Can be plural or singular (*book*, *books*)

Can have determiners (*the book*)

Can be modified by adjectives (*blue book*)

Can have possessors (*my books*, *John's book*)

ENGLISH VERBS

Take noun phrases as arguments

- At least a subject
    - *Dr. A parsed aggressively.*
- *Sometimes one or two objects*
    - *Dr. A parsed the data.*
    - *Dr. A passed [the function] [an argument].*

Can take tense morphology (past/non-past)

Can be modified by adverbs

ENGLISH ADJECTIVES

Modify nouns (restrict their reference)

- *his heavy book.*
- *His book is heavy.*

Can take comparative/superlative suffixes when allowed by prosody

- *heavy, heavier, heaviest*
- *But pitiful, more pitiful, most pitiful*

Not all languages have adjectives

Some languages (e.g. Korean, Hmong, Vietnamese) use verbs to modify nouns in this way

ENGLISH ADVERBS

Modify verbs, adjectives, and other adverbs (restrict their reference)

- *He erroneously concluded that PHP is a real programming language simply because it is Turing complete.*
- *The design of PHP is exceptionally poor.*

CLOSED CLASS PARTS OF SPEECH

ENGLISH PREPOSITIONS

Occur before noun phrases

Relate noun phrase to some higher-level constituent

- *He lingered in the depths of despair.*

It is not difficult to characterize prepositions **formally,** but they are very difficult to characterize **semantically**
(a good argument to not introduce semantic considerations into PoS categories)

They are often identical in spelling and pronunciation to **particles.**

ENGLISH DETERMINERS

Determiners are words that come at the beginning of English noun phrases

Articles like *the*, *a*, and *an*

- *The interpreter choked on an unknown identifier.*

Other determiners include some demonstratives like *this* and *that*.

- *That version of Python really chaps my hide.*

ENGLISH PRONOUNS

Pronouns can replace noun phrases, acting as a short hand for them

- *You code like a wizard.*
- *Who knows Haskell, really?.*

ENGLISH CONJUNCTIONS

Conjunctions join phrases, clauses, or sentences.

Typically, the conjuncts joined by a conjunction are of the same type

Coordinating conjunctions:

- *and, or, but, …*

Subordinating conjunctions:

- *if, because, though, while, …*

ENGLISH AUXILIARY VERBS

"Helping verbs" that occur before main verbs

Some can occur as main verbs as well

- *Be*
    - *I am the type system. (main)*
    - *I am working on my project. (aux verb)*
- *Have*

Others (e.g. modals) occur only as auxiliary verbs

- *must, might, would, will, could, can, …*

ENGLISH PARTICLES

Particle is sometimes used as a grab-bag category for closed-class items that do not fir in other categories

Most often, in English, these resemble prepositions or adverbs and are used in combination with a verb

- *He tore off his shirt.*
- *He tore his shirt off.*
- *I want to leave.*

NUMERALS

Numerals have properties of both nouns and adjectives

They can by the subject and object of verbs:

- *Two will enter but only one will leave.*
- *I bought twenty.*

They can function both attributively and predicatively:

- *Two variables where left undeclared.*
- *We are three.*

When used attributively, they come before any adjectives:

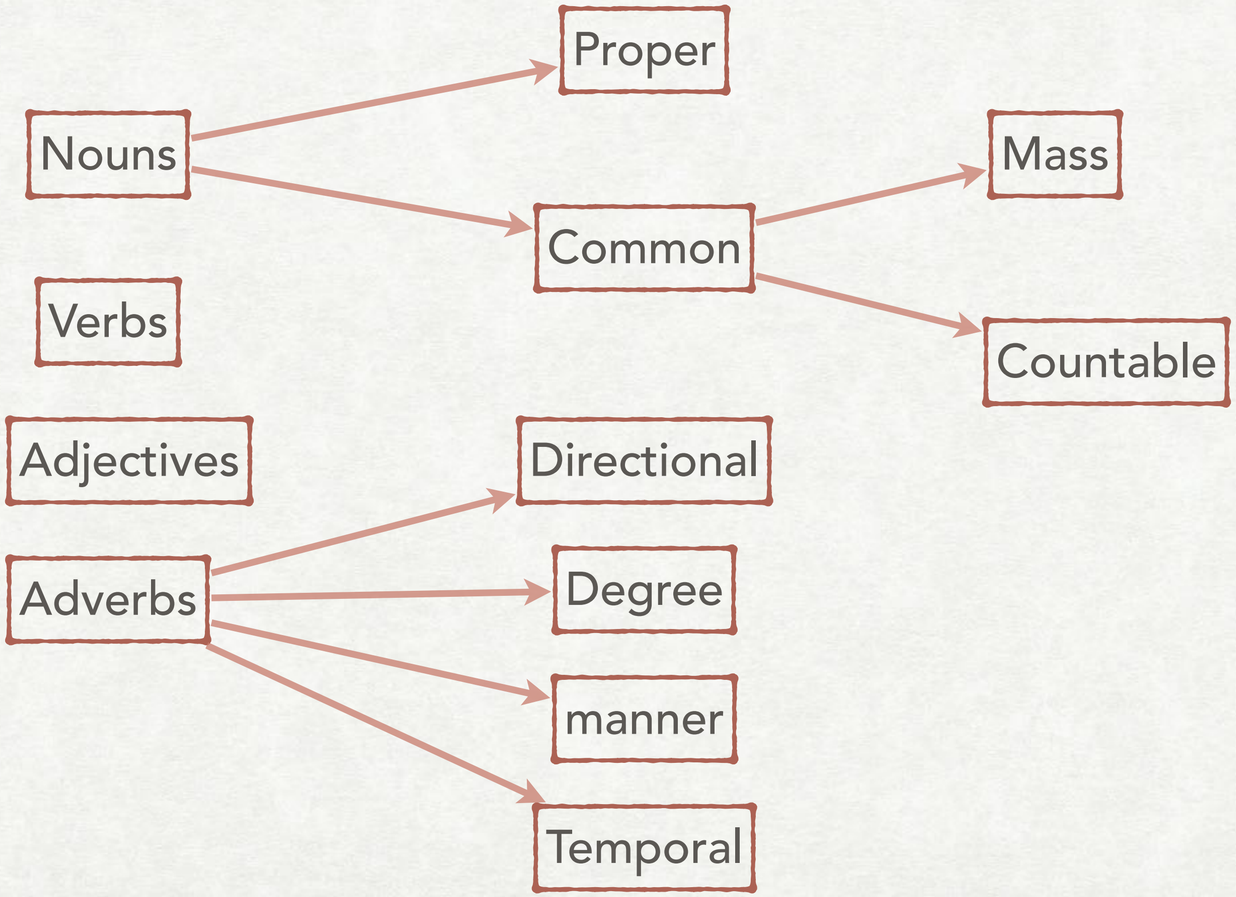
- *The two undeclared variables where the cause of much consternation.*
- **The undeclared two variables where the cause of much consternation.*

HOW DO WE KNOW THE CLASS?

1. Substitution test

The ADJ cat sat on the mat
The blue NOUN sits on the NOUN
The blue cat VERB on the mat
The blue cat sat PREP the mat

FINE-GRAINED CLASSES



HARD CASES

I will call up my friend

I will call my friend up

I will call my friend up in the treehouse

Gerunds

- I like walking
- I like apples
- His walking kept him fit
- His apples kept him fit
- His walking slowly kept him fit
- His apples slowly kept him fit

Potential Additional Classes

Interjections
Negatives
Politeness markers
Greetings
Existential *there*
Numbers, Symbols, Money, ….
Emojis
URLs
Hashtags

GOOGLE'S UNIVERSAL TAGS

ADJ: adjective
ADP: adposition (preposition or postposition)
ADV: adverb
AUX: auxiliary
CCONJ: coordinating conjunction
DET: determiner
INTJ: interjection
NOUN: noun
NUM: numeral
PART: particle
PRON: pronoun
PROPN: proper noun
PUNCT: punctuation
SCONJ: subordinating conjunction
SYM: symbol
VERB: verb
X: other

Warning

Don't tell a linguist these are *truly* universal
They will be very offended — and they will be right to do so.

But, they can be very useful!

WHY DO WE NEED MODELS? IS POS TAGGING HARD?

If every "word" could only be associated with a single tag, PoS tagging would be trivial

- How would you do it?
- Do you foresee any problems?

But, this won't always work

- *lives* can be a noun or a verb
- *black* can be an adjective, verb, proper noun, common noun, etc…

How bad is this problem, really?

PoS tags per orthographic word in Penn Treebank

7 down	5 out
6 that	5 many
6 set	5 less
6 put	5 left
6 open	5 Japanese
6 hurt	5 in
6 cut	5 hit
6 bet	5 half
6 back	5 further
5 vs.	5 forecast
5 the	5 fit
5 spread	5 first
5 split	5 East
5 say	5 counter
5 's	5 cost
5 run	5 close
5 repurchase	5 bid
5 read	5 beat
5 present	5 a

317 down RB
200 down RP
138 down IN
10 down JJ
1 down VBP
1 down RBR
1 down NN

MODELING APPROACH

1. Pick the most frequent tag per word
   - What accuracy do you think it would give you over average English text?

2. Look at the context
   - Preceding (and succeeding) words
   - Preceding (and succeeding) tags
   - The …
   - To …
   - John’s blue …

THE OUT-OF-VOCABULARY PROBLEM

How do you handle cases where your dictionary does not include all of the words?

Proper names?
Borrowed words?
Neologisms?

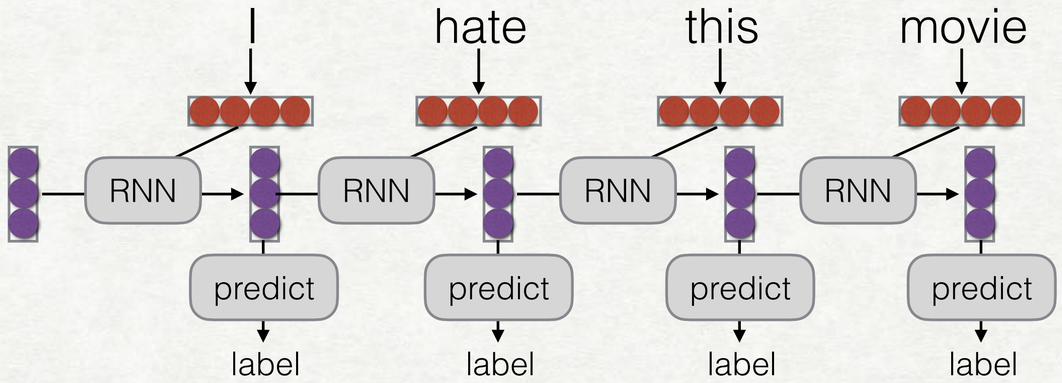
As a language user, these are not a problem for you.

How would you give a POS-tagger the same superpower?

Stay tuned for the CRF class!

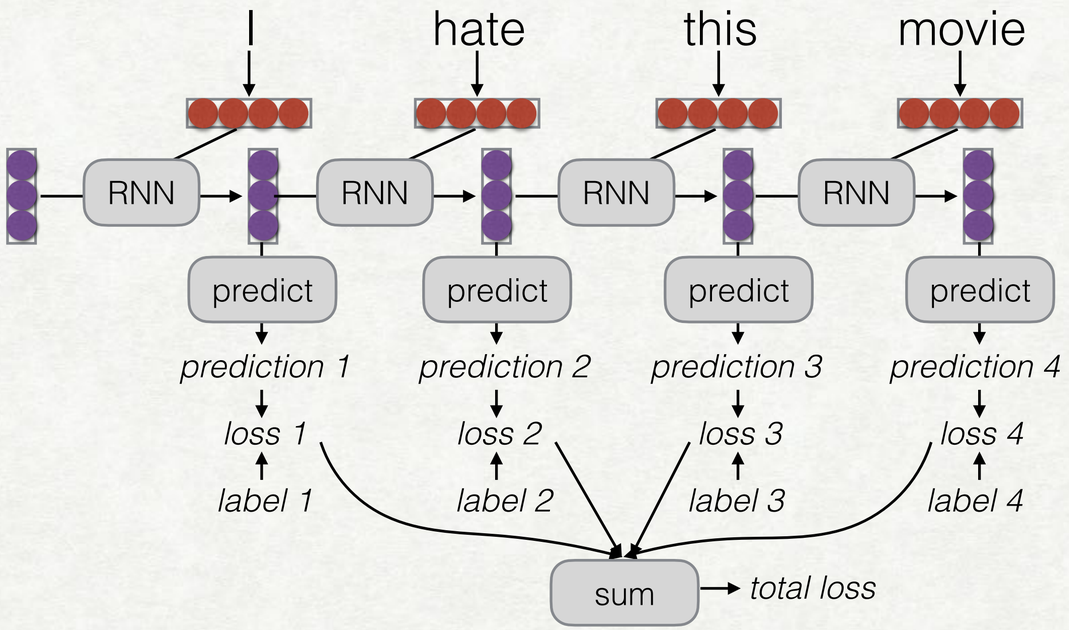
NEURAL POS TAGGING

TRAINING A NEURAL POS TAGGER



TRAINING A NEURAL POS TAGGER

Calculating the loss



CODE EXAMPLE

NEXT CLASS PREVIEW

Classification 101