February 26, 2018 at 1:18 PM

CIS 530 - 2/26/18

Announcements:

1. No quiz on Wednesday. I have decided to give you that time so that you can study for midterms in your other classes.
2. Wednesday's class will be a lecture rather than an in-class exercise. I'm thinking through what changes to make to the class based on your mid-semester feedback. Thank you for your thoughtful suggestions!
3. Instead of a programming assignment this week, we will have you do the first part of the midterm project. Details:
   - Form a team [min size 4, max 6]
   - Come up with 3 project ideas
   - The goal of the term project is to design a homework assignment similar to ones you have completed in class.
   - Your project will consist of these components:
     1) A description of the problem, in the style of the homework write-ups
     2) Training and evaluation data
     3) A commented implementation of the simplest baseline that solves the problem
     4) " " " of a baseline published in the literature
     5) One extension per team member that attempts to beat the baseline

Shared tasks
- SEMEval
- CoNLL
- WMT

- Your project will consist of these components

Shared tasks
- SEMEval
- CoNLL
- WMT

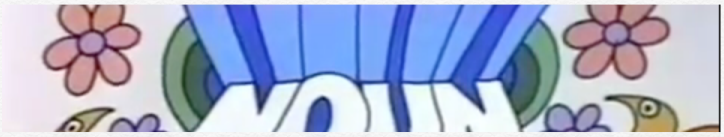
1) A description of the problem, in the style of the homework write-ups
2) Training and evaluation data
3) A commented implementation of the simplest baseline that solves the problem
4) " " " of a baseline published in the literature
5) One extension per team member that attempts to beat the baseline.
6) A scoring script implementing an objective function that can be used to score submissions on the leaderboard

NLG: BLEU Rouge

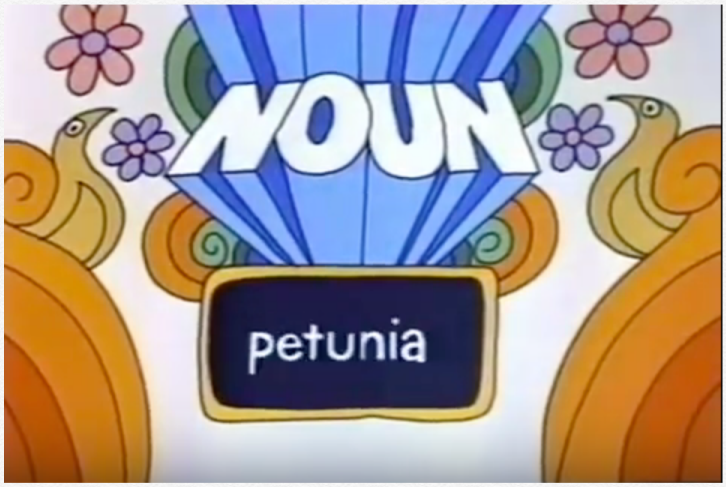
- These will be delivered in different milestones
- The first milestone is due Wednesday after spring break. (March 14)
  1) Form a team
  2) Submit a short ~1-2 paragraph description of your 3 ideas.

We will use class on Monday after Spring Break to help you.

Todays lecture: Part of Speech Tagging (Chapter 10 in the textbook)



Todays lecture: Part of Speech Tagging (Chapter 10 in the textbook)



Ancient Greek tag set: (c. 100 BC)
noun, verb, pronoun, preposition, adverb, conjunction, participle, article

Schoolhouse Rock tag set: (c. 1970)
+ adjective + interjection
- participle - article

Every word in Vocbulary belongs to one or more of these word classes

Assigning the classes to words in a sentence is called POS tagging.

Many words can have multiple POS tags. Can you think of some?

Four major classes ← open classes

Nouns Verbs Adjectives Adverbs

English has all 4 but not every language does

NOUN - person, place, or thing

↳ Proper Nouns names of specific entities or people Prof. Marcus University of Pennsylvania NRA

↳ common nouns

↳ count nouns (allow grammatical enumeration; occuring in both singular and plural ungrammatical * cat is cute — cats are cute

↳ mass nouns (conceptualized as homogenous groups, cannot be pluralized, can appear without determiners even in singular form)

the cats
a cereal
democracy

snow
salt
communism

grammatical the snow is white — snow is white

Verbs - words describing actions and processes

English verbs have inflectional markers.

- 3rd person singular
- non- " " "
- progressive (ing)
- past

compute	
it computes	+s
they/you/I compute	
computing	+ing
computed	+ed

Adjectives - words that describe properties or qualities

Color, age, value, good/bad material, subjective words

Scalar adjectives
temp: lukewarm < warm < hot
boiling < scalding <

Adjectives — words that describe properties or qualities
- color, age, value, material, subjective words (good/bad, amazing)
- scalar adjectives — temp: lukewarm < warm < hot < boiling < scalding < face-melting

Adverbs — modify verbs or whole verb phrases, or other words like adjectives
- locatives: here, home, uphill
- degree: very, extremely, extraordinarially, somewhat, not really, -ish
  - very warm → hot
  - very delicious → scrumptious
- manner: slowly/quickly, softly/gently, alluringly
- temporal: yesterday, Monday, last semester

Other POS tags in English

numerals: one, two, nth, fist, second,

closed classes:
- prepositions: of, on, over, under, to, from, around — "anywhere a hippo can go"
- determiners: some, a, an (indefinite); the, this, that (definite) — Español: el-masc., la-fem, los pl+m, las pl+fem
- pronouns: she, he, it, they, who, whoever, whatever.
- conjunctions: and, or, but — I really like the book that Sally wrote.
- auxiliary verbs: was hacked — when tense, passive voice, aspect ← still going on?, polarity ← negation, mood ← desirable/possible/suggested
- particles: prepositions joined to a verb

Tag	Description	Example	Tag	Description	Example
CC	coordin. conjunction	*and, but, or*	SYM	symbol	+,%, &
CD	cardinal number	*one, two*	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, sing.	*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*			

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

There/EX are/VBP 70/CD children/NNS **there/RB**

Preliminary/JJ findings/NNS were/VBD **reported/VBN** in/IN today/NN **'s/POS** New/NNP England/NNP Journal/NNP of/IN Medicine/NNP ./.

Corpora with manual POS tags:

- Brown corpus – 1 million words of 500 written English texts from different genres
- WSJ corpus – 1 million words from the Wall Street Journal

Corpora with manual POS tags:

- Brown corpus – 1 million words of 500 written English texts from different genres
- WSJ corpus – 1 million words from the Wall Street Journal
- Switchboard corpus – 2 M words of telephone conversations

POS Tagging

words are ambiguous

so tagging must resolve / disambiguate

Types:		WSJ	Brown
Unambiguous	(1 tag)	44,432 (**86%**)	45,799 (**85%**)
Ambiguous	(2+ tags)	7,025 (**14%**)	8,050 (**15%**)
Tokens:			
Unambiguous	(1 tag)	577,421 (**45%**)	384,349 (**33%**)
Ambiguous	(2+ tags)	711,780 (**55%**)	786,646 (**67%**)

Figure 10.2 The amount of tag ambiguity for word types in the Brown and WSJ corpora, from the Treebank-3 (45-tag) tagging. These statistics include punctuation as words, and assume words are kept in their original case.

Some words have up to 6 tags