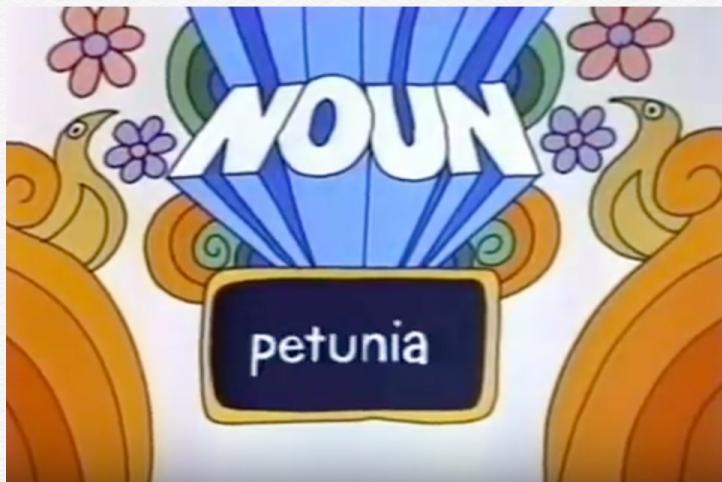


< Notes



Today's lecture: Part of Speech Tagging
(Chapter 10 in the textbook)



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

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

Every word in Vocabulary 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?

< Notes



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 NRA University of Pennsylvania

common nouns

the cats
a cereal
democracy

snow
salt
communism

count nouns (allow grammatical enumeration, occurring 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 snow is white
grammatical 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

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

- words that describe qualities

Properties or good/bad
color, age, value,
material, subjective words



< Notes



Adjectives - words that describe properties or qualities like 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

very
warm
→
hot

↳ locatives: here, home, uphill

very
delicious
→
scrumptious

↳ degree: very, extremely, extraordinarily, somewhat, not really, ish.

↳ Manner: slowly / quickly, softly / gently, alluringly.

↳ temporal: yesterday, Monday, last semester

Other POS tags in English

numerals: one, two, nth, first, second,

prepositions: of, on, over, under, to, from, around, "anywhere a hippo can go" Espanol el-marc.
determiners: some, a, an, the, this, that la-fun los pl+ las pt+ fem
Indefinite definite

pronouns: she, he, it, they who whenever whatever.

closed classes

conjunctions: and or but
I really like the book that Sally wrote. when

auxiliary verbs: was hacked tense → passive voice

particles: preposition joins aspect → still going on?
↓ a verb polarity → negation mood → desirable / possible / suggested

< Notes



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

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



Notes



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



< Notes



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

- 1 earnings took a back/Adj seat
- 2 a small yard in the back/noun
- 3 senators back/<sup>present
tense verb</sup> the bill
- 4 He started to back/<sup>initial
verb</sup> toward the door
- 5 to buy back/^{particle} stock
- 6 I was young back/<sup>Adverb
temporal</sup> then



HMMs for tagging

Hidden Markov Models are trained on tagged corpora; instead of using the Baum-Welch algorithm. Fully supervised instead, with MLE parameter estimation

Basic eqn. for HMM tagging:

$$\hat{t}_i^n = \arg \max_{t_i^n} P(t_i^n | w_i^n)$$

Use Bayes Rule

$$= \arg \max_{t_i^n} \frac{P(w_i^n | t_i^n) P(t_i^n)}{P(w_i^n)}$$

$$= \arg \max_{t_i^n} P(w_i^n | t_i^n) P(t_i^n)$$

⑤ Simplifying assumptions



< Notes



Simplifying assumptions:

- ① Probability of a word only depends on its own tag, and it is independent of neighboring words and tags:

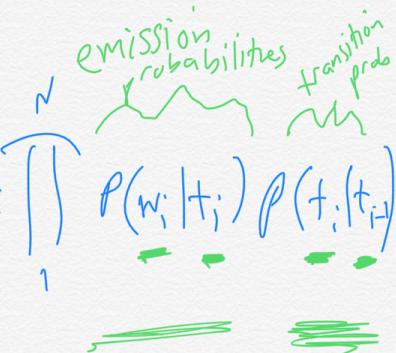
$$p(w_i^n | t_i^n) \approx \prod_{i=1}^N p(w_i | t_i)$$

- ② Bigram assumption. The probability of a tag depends only on prev. tag, not the whole tag sequence:

$$\underline{p(t_i^n)} \approx \prod_{i=1}^N \underline{p(t_i | t_{i-1})}$$

Combining everything

$$\hat{t}_i^n = \operatorname{argmax}_{t_i^n} P(t_i^n | w_i^n) \approx \operatorname{argmax}_{t_i^n} \prod_{i=1}^N p(w_i | t_i) p(t_i | t_{i-1})$$



< Notes



Estimating probabilities

Tag transition probabilities

MLE

$$P(t_i | t_{i-1}) = \frac{\text{Count}(t_{i-1}, t_i)}{\text{Count}(t_{i-1})}$$

Example

$$P(VB | MD) = \frac{\text{Count}(MD, VB)}{\text{Count}(MD)} = \frac{10471}{13124} = .8$$

Emission probabilities : probability that given a tag, we will see a word

$$P(w_i | t_i) = \frac{\text{Count}(t_i, w_i)}{\text{Count}(t_i)}$$

Example

$$P(\text{will} | MD) = \frac{\text{Count}(\text{will}, MD)}{\text{Count}(MD)}$$

$$= \frac{4061}{13124}$$

$$= .31$$

(house) (will)



< Notes



Extending HMMs to Trigrams

use more word history

$$P(t_i^n) \approx \prod_{i=1}^n P(t_i | t_{i-1}, t_{i-2})$$

estimation = MLE

Advanced features:

① Interpolation (just like LMs)

λ_3 Trigram tags $p(t_i | t_{i-1}, t_{i-2}) +$

λ_2 bigrams $p(t_i | t_{i-1}) +$

λ_1 unigrams $p(t_i) = \frac{\text{count}(t_i)}{N}$

② Dealing with Unknown words

Many new words are added to the lexicon

Most of our annotated corpora are 20+ years old.

Never contained the word, google



< Notes



- Morphological clues
 - wug(s) → plural noun
 - ed → past tense verb
 - able → adjective
 - ly → adverb
- Capitalization

- Tag Sequence

$$p(t_i | t_{i-1} t_{i-2})$$

Beyond HMMs: MEMM
 Maximum entropy Markov Models

- discriminative sequence model

Instead of

$$\hat{T} = \underset{T}{\operatorname{arg\,max}} \underbrace{P(T|w)}_{\rightarrow} = \underset{T}{\operatorname{arg\,max}} P(w|T) P(T)$$



Notes

Instead of

$$\hat{T} = \arg \max_T P(T | w)$$

$$\rightarrow = \arg \max_T P(w | T) P(T)$$

$$\rightarrow = \arg \max_T \prod_i P(w_i | t_i) \prod_i P(t_i | t_{i-1})$$

MEMMs compute posterior directly, and train it to discriminate among tag sequences

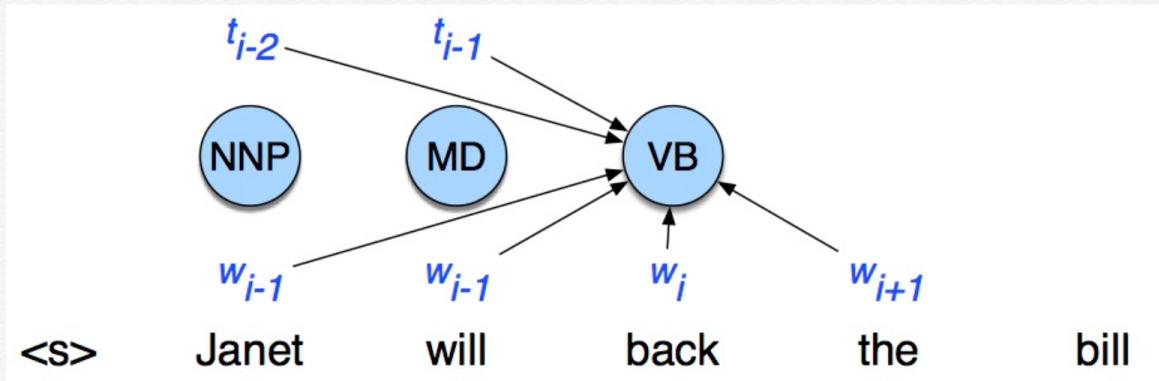
$$\hat{T} = \arg \max_T P(T | w)$$

$$= \arg \max_T \prod_i P(t_i | w_i, t_{i-1})$$

or use tons of features



< Notes



$$\begin{aligned}
 &= \underset{T}{\operatorname{argmax}} \prod_i p(t_i | w_{i-1}^{i+1}, t_{i-1}^{i+1}) \\
 &= \underset{T}{\operatorname{argmax}} \frac{\exp \left(\sum_i w_i f_i(t_i, \dots) \right)}{\sum_{t' \in \text{tagset}} \exp \left(\sum_i w_i f_i(t', \dots) \right)}
 \end{aligned}$$

What does this remind you of?
logistic regression

< Notes



What does this remind you of?
logistic regression

Most Frequent class baseline

92% accuracy

versus

97% SOA taggers

Q: Does that mean that POS tagging
is "solved"?

Notes



ikr smh he asked fir yo last name so he can add u on fb
lololol

:o :/ :(>:o (: :) >.< XD -__-
o.O ;D :-@ @_ :P 8D :1 >:(:D =|
") >

<u>word</u>	<u>tag</u>	<u>confidence</u>
ikr	!	0.8143
smh	G	0.9406
he	O	0.9963
asked	V	0.9979
fir	P	0.5545
yo	D	0.6272
last	A	0.9871
name	N	0.9998
so	P	0.9838
he	O	0.9981
can	V	0.9997
add	V	0.9997
u	O	0.9978
on	P	0.9426
fb	^	0.9453
lololol	!	0.9664

<u>word</u>	<u>tag</u>	<u>confidence</u>
:o	E	0.9387
:/	E	0.9983
:(E	0.9975
>:o	E	0.9994
(:	E	0.9994
):	E	0.9997
>.<	E	0.9952
XD	E	0.9938
-__-	E	0.9956
o.O	E	0.9899
;D	E	0.9995
:-)	E	0.9992
@_@	E	0.9964
:P	E	0.9996
8D	E	0.9961
:	E	0.6925
1	\$	0.9194
>:(E	0.9715
:D	E	0.9996
=	E	0.9963
"	,	0.6125
)	,	0.9078
:	,	0.7460
>	G	0.7490
...	,	0.5223
.	,	0.9946

- "ikr" means "I know, right?", tagged as an interjection.
- "so" is being used as a subordinating conjunction, which our coarse tagset denotes *P*.
- "fb" means "Facebook", a very common proper noun (*A*).
- "yo" is being used as equivalent to "your"; our coarse tagset has possessive pronouns as *D*.
- "fir" is a misspelling or spelling variant of the preposition *for*.
- Perhaps the only debatable errors in this example are for *ikr* and *smh* ("shake my head"): should they be *G* for miscellaneous acronym, or *!* for interjection?

Challenge case for emoticon segmentation/recognition: 20/26 precision, 18/21 recall.

Most POS tagging work is evaluated on the WSJ.
This is a well-edited formally written newspaper
that obeys standard conventions. Twitter OTOH is...

- Conversational
- lacks conventional spelling - confe!e!
- introduces abbreviations and other innovations to match the 140 character limit (now raised to 280).

CMU researchers introduced a new tagset for
L...ll or pl...c some manually annotated Tweets.



< Notes



The performance of their system is in the mid-80% range instead of high 90s.

Other challenges

Other languages

Low resource languages

Dialects.

Q: Do all language share the same parts of speech?

Language	Source	# Tags	O/O	U/U	O/U
Arabic	PADT/CoNLL07 (Hajič et al., 2004)	21	96.1	96.9	97.0
Basque	Basque3LB/CoNLL07 (Aduriz et al., 2003)	64	89.3	93.7	93.7
Bulgarian	BTB/CoNLL06 (Simov et al., 2002)	54	95.7	97.5	97.8
Catalan	CESS-ECE/CoNLL07 (Martí et al., 2007)	54	98.5	98.2	98.8
Chinese	Penn ChineseTreebank 6.0 (Palmer et al., 2007)	34	91.7	93.4	94.1
Chinese	Sinica/CoNLL07 (Chen et al., 2003)	294	87.5	91.8	92.6
Czech	PDT/CoNLL07 (Böhmová et al., 2003)	63	99.1	99.1	99.1
Danish	DDT/CoNLL06 (Kromann et al., 2003)	25	96.2	96.4	96.9
Dutch	Alpino/CoNLL06 (Van der Beek et al., 2002)	12	93.0	95.0	95.0
English	PennTreebank (Marcus et al., 1993)	45	96.7	96.8	97.7
French	FrenchTreebank (Abeillé et al., 2003)	30	96.6	96.7	97.3
German	Tiger/CoNLL06 (Brants et al., 2002)	54	97.9	98.1	98.8
German	Negra (Skut et al., 1997)	54	96.9	97.9	98.6
Greek	GDT/CoNLL07 (Prokopidis et al., 2005)	38	97.2	97.5	97.8
		12	91.7	95.6	95.6
		28	94.9	95.8	95.8



< Notes



Q: Do all language share the same parts of speech?

Language	Source	# Tags	O/O	U/U	O/U
Arabic	PADT/CoNLL07 (Hajič et al., 2004)	21	96.1	96.9	97.0
Basque	Basque3LB/CoNLL07 (Aduriz et al., 2003)	64	89.3	93.7	93.7
Bulgarian	BTB/CoNLL06 (Simov et al., 2002)	54	95.7	97.5	97.8
Catalan	CESS-ECE/CoNLL07 (Martí et al., 2007)	54	98.5	98.2	98.8
Chinese	Penn ChineseTreebank 6.0 (Palmer et al., 2007)	34	91.7	93.4	94.1
Chinese	Sinica/CoNLL07 (Chen et al., 2003)	294	87.5	91.8	92.6
Czech	PDT/CoNLL07 (Böhmová et al., 2003)	63	99.1	99.1	99.1
Danish	DDT/CoNLL06 (Kromann et al., 2003)	25	96.2	96.4	96.9
Dutch	Alpino/CoNLL06 (Van der Beek et al., 2002)	12	93.0	95.0	95.0
English	PennTreebank (Marcus et al., 1993)	45	96.7	96.8	97.7
French	FrenchTreebank (Abeillé et al., 2003)	30	96.6	96.7	97.3
German	Tiger/CoNLL06 (Brants et al., 2002)	54	97.9	98.1	98.8
German	Negra (Skut et al., 1997)	54	96.9	97.9	98.6
Greek	GDT/CoNLL07 (Prokopidis et al., 2005)	38	97.2	97.5	97.8
Hungarian	Szeged/CoNLL07 (Csendes et al., 2005)	43	94.5	95.6	95.8
Italian	ISST/CoNLL07 (Montemagni et al., 2003)	28	94.9	95.8	95.8
Japanese	Verbmobil/CoNLL06 (Kawata and Bartels, 2000)	80	98.3	98.0	99.1
Japanese	Kyoto4.0 (Kurohashi and Nagao, 1997)	42	97.4	98.7	99.3
Korean	Sejong (http://www.sejong.or.kr)	187	96.5	97.5	98.4
Portuguese	Floresta Sintá(c)tica/CoNLL06 (Afonso et al., 2002)	22	96.9	96.8	97.4
Russian	SynTagRus-RNC (Boguslavsky et al., 2002)	11	96.8	96.8	96.8
Slovene	SDT/CoNLL06 (Džeroski et al., 2006)	29	94.7	94.6	95.3
Spanish	Ancora-Cast3LB/CoNLL06 (Civit and Martí, 2004)	47	96.3	96.3	96.9
Swedish	Talbanken05/CoNLL06 (Nivre et al., 2006)	41	93.6	94.7	95.1
Turkish	METU-Sabancı/CoNLL07 (Oflazer et al., 2003)	31	87.5	89.1	90.2

Does it make sense to have a "universal" pos tag set?



< Notes



A Universal Part-of-Speech Tagset

Slav Petrov
 Google Research
 New York, NY, USA
 slav@google.com

Dipanjan Das
 Carnegie Mellon University
 Pittsburgh, PA, USA
 dipanjan@cs.cmu.edu

Ryan McDonald
 Google Research
 New York, NY, USA
 ryanmcd@google.com

Abstract

To facilitate future research in unsupervised induction of syntactic structure and to standardize best-practices, we propose a tagset that consists of twelve universal part-of-speech categories. In addition to the tagset, we develop a mapping from 25 different treebank tagsets to this universal set. As a result, when combined with the original treebank data, this universal tagset and mapping produce a dataset consisting of common parts-of-speech for 22 different languages. We highlight the use of this resource via two experiments, including one that reports competitive accuracies for unsupervised grammar induction without gold standard part-of-speech tags.

forms across languages. These categories are often called *universals* to represent their cross-lingual nature (Carnie, 2002; Newmeyer, 2005). For example, Naseem et al. (2009) used the Multext-East (Erjavec, 2004) corpus to evaluate their multi-lingual POS induction system, because it uses the same tagset for multiple languages. When corpora with common tagsets are unavailable, a standard approach is to manually define a mapping from language and treebank specific fine-grained tagsets to a predefined universal set. This was the approach taken by Das and Petrov (2011) to evaluate their cross-lingual POS projection system for six different languages.

To facilitate future research and to standardize best-practices, we propose a tagset that consists of twelve universal POS categories. While there

sentence:	The	oboist	Heinz	Holliger	has	taken	a	hard	line	about	the	problems	.
original:	DT	NN	NNP	NNP	VBZ	VBN	DT	JJ	NN	IN	DT	NNS	.
universal:	DET	NOUN	NOUN	NOUN	VERB	VERB	DET	ADJ	NOUN	ADP	DET	NOUN	.

Figure 1: Example English sentence with its language specific and corresponding universal POS tags.



inflectional

morphology

pluralization





inflectional morphology - pluralization
tense

derivational morphology - oboe
→ oboist "one ~~verb~~
piano verbs
→ pianist the noun"

agglutinative languages

words contain a lot of morphemes joined together, which would be expressed with separate words in other languages