Word Classes and Part-of-Speech (POS) Tagging

Based on slides from Julia Hirschberg -
www.cs.columbia.edu/~julia

Garden Path Sentences

- The old dog

……the footsteps of the young.

- The cotton clothing

……is made of grows in Mississippi.

- The horse raced past the barn

……fell.

Word Classes

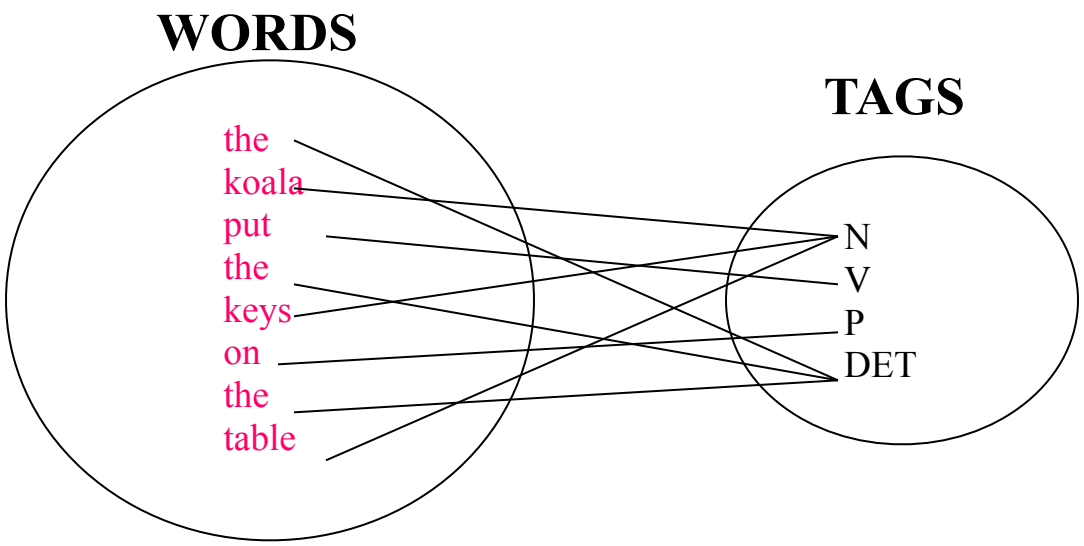
- Words that somehow 'behave' alike:
  - Appear in similar contexts
  - Perform similar functions in sentences
  - Undergo similar transformations
- ~9 traditional word classes of parts of speech
  - Noun, verb, adjective, preposition, adverb, article, interjection, pronoun, conjunction

Some Examples

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

Defining POS Tagging

- The process of assigning a part-of-speech or lexical class marker to each word in a corpus:



Applications for POS Tagging

- Speech synthesis pronunciation
  - *Lead* *Lead*
  - *INsult* *inSULT*
  - *OBject* *obJECT*
  - *OVERflow* *overFLOW*
  - *DIScount* *disCOUNT*
  - *CONtent* *conTENT*
- Parsing: e.g. *Time flies like an arrow*
  - Is *flies* an N or V?
- Word prediction in speech recognition
  - Possessive pronouns (*my, your, her*) are likely to be followed by nouns
  - Personal pronouns (*I, you, he*) are likely to be followed by verbs
- Machine Translation

Closed vs. Open Class Words

- Closed class: relatively fixed set
  - Prepositions: of, in, by, …
  - Auxiliaries: may, can, will, had, been, …
  - Pronouns: I, you, she, mine, his, them, …
  - Usually function words (short common words which play a role in grammar)
- Open class: productive
  - English has 4: Nouns, Verbs, Adjectives, Adverbs
  - Many languages have all 4, but not all!
  - In Lakhota and possibly Chinese, what English treats as adjectives act more like verbs.

Open Class Words

- Nouns
  - Proper nouns
    - Harvey Mudd College, Claremont, Zach Dodds, Los Angeles County Museum of Art
    - English capitalizes these
    - Many have abbreviations
  - Common nouns
    - All the rest
    - German capitalizes these.

- Count nouns vs. mass nouns
  - Count: Have plurals, countable: goat/goats, one goat, two goats
  - Mass: ***Not*** countable (fish, salt, communism) (?two fishes)
- Adjectives: identify properties or qualities of nouns
  - Color, size, age, …
  - Adjective ordering restrictions in English:
    - Old blue book, ***not*** Blue old book
  - In Korean, adjectives are realized as verbs
- Adverbs: also modifiers (of verbs, adjectives, adverbs)
  - The very happy man walked home extremely slowly yesterday.

  - Directional/locative adverbs (here, home, downhill)
  - Degree adverbs (extremely, very, somewhat)
  - Manner adverbs (slowly, slinkily, delicately)
  - Temporal adverbs (Monday, tomorrow)
- Verbs:
  - In English, take morphological affixes (eat/eats/eaten)
  - Represent actions (walk, ate), processes (provide, see), and states (be, seem)
  - Many subclasses, e.g.
    - eats/V ⇒ eat/VB, eat/VBP, eats/VBZ, ate/VBD, eaten/VBN, eating/VBG, ...
    - Reflect morphological form & syntactic function

How Do We Assign Words to Open or Closed?

- Nouns denote people, places and things and can be preceded by articles? But…

  My ***typing*** is very bad.

  *The ***Mary*** loves John.

- Verbs are used to refer to actions, processes, states
  - But some are closed class and some are open

  I ***will have emailed*** everyone by noon.

- Adverbs modify actions
  - Is Monday a temporal adverbial or a noun?

Closed Class Words

- Idiosyncratic
- Closed class words (Prep, Det, Pron, Conj, Aux, Part, Num) are generally easy to process, since we can enumerate them….but
  - Is it a Particles or a Preposition?
    - George eats up his dinner/George eats his dinner up.
    - George eats up the street/*George eats the street up.
  - Articles come in 2 flavors: definite (the) and indefinite (a, an)
    - What is this in 'this guy…'?

Choosing a POS Tagset

- To do POS tagging, first need to choose a set of tags
- Could pick very coarse (small) tagsets
  – N, V, Adj, Adv.
- More commonly used: Brown Corpus (Francis & Kucera ‘82), 1M words, 87 tags – more informative but more difficult to tag
- Most commonly used: Penn Treebank: hand-annotated corpus of *Wall Street Journal*, 1M words, 45-46 subset

Penn Treebank Tagset

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 Penn Treebank Tags

- The/DT grand/JJ jury/NN commmented/VBD on/IN a/DT number/NN of/IN other/JJ topics/NNS ./.
- Prepositions and subordinating conjunctions marked IN (“although/IN I/PRP..”)
- Except the preposition/complementizer “to” is just marked “TO”

Tag Ambiguity

- Words often have more than one POS: *back*
  - The *back* door = JJ
  - On my *back* = NN
  - Win the voters *back* = RB
  - Promised to *back* the bill = VB
- The POS tagging problem is ***to determine the POS tag for a particular instance of a word***

Tagging Whole Sentences with POS is Hard

- Ambiguous POS contexts
  - E.g., Time flies like an arrow.
- Possible POS assignments
  - Time/[V,N] flies/[V,N] like/[V,Prep] an/Det arrow/N
  - Time/N flies/V like/Prep an/Det arrow/N
  - Time/V flies/N like/Prep an/Det arrow/N
  - Time/N flies/N like/V an/Det arrow/N
  - …..

How Big is this Ambiguity Problem?

		Original 87-tag corpus		Treebank 45-tag corpus	
Unambiguous (1 tag)		**44,019**		**38,857**	
Ambiguous (2–7 tags)		**5,490**		**8844**	
Details:	2 tags	4,967		6,731	
	3 tags	411		1621	
	4 tags	91		357	
	5 tags	17		90	
	6 tags	2	(*well*, *beat*)	32	
	7 tags	2	(*still*, *down*)	6	(*well*, *set*, *round*, *open*, *fit*, *down*)
	8 tags			4	(*'s*, *half*, *back*, *a*)
	9 tags			3	(*that*, *more*, *in*)

How Do We Disambiguate POS?

- Many words have only one POS tag (e.g. is, Mary, very, smallest)
- Others have a single ***most likely*** tag (e.g. a, dog)
- Tags also tend to *co-occur* regularly with other tags (e.g. Det, N)
- In addition to conditional probabilities of words $P(w_1|w_{n-1})$, we can look at POS likelihoods ($P(t_1|t_{n-1})$) to disambiguate sentences and to assess sentence likelihoods

Some Ways to do POS Tagging

- Rule-based tagging
  - E.g. EnCG ENGTWOL tagger
- Transformation-based tagging
  - Learned rules (statistical and linguistic)
  - E.g., Brill tagger
- Stochastic, or, Probabilistic tagging
  - HMM (Hidden Markov Model) tagging

Rule-Based Tagging

- Typically…start with a dictionary of words and possible tags
- Assign all possible tags to words using the dictionary
- Write rules by hand to *selectively remove* tags
- Stop when each word has exactly one (presumably correct) tag

Start with a POS Dictionary

- she: PRP
- promised: VBN,VBD
- to: TO
- back: VB, JJ, RB, NN
- the: DT
- bill: NN, VB
- Etc… for the ~100,000 words of English

Assign All Possible POS to Each Word

			NN		
			RB		
	VBN		JJ		VB
PRP	VBD	TO	VB	DT	NN
She	**promised**	**to**	**back**	**the**	**bill**

Apply Rules Eliminating Some POS

E.g., *Eliminate VBN if VBD is an option when VBN/VBD follows "<start> PRP"*

			NN		
			RB		
	VBN		JJ		VB
PRP	VBD	TO	VB	DT	NN
She	**promised**	**to**	**back**	**the**	**bill**

Apply Rules Eliminating Some POS

E.g., *Eliminate VBN if VBD is an option when VBN/VBD follows "<start> PRP"*

			NN		
			RB		
			JJ		VB
PRP	VBD	TO	VB	DT	NN
She	**promised**	**to**	**back**	**the**	**bill**

EngCG ENGTWOL Tagger

- Richer dictionary includes morphological and syntactic features (e.g. subcategorization frames) as well as possible POS
- Uses two-level morphological analysis on input and returns all possible POS
- Apply negative constraints (> 3744) to rule out incorrect POS

Sample ENGTWOL Dictionary

Word	POS	Additional POS features
smaller	ADJ	COMPARATIVE
entire	ADJ	ABSOLUTE ATTRIBUTIVE
fast	ADV	SUPERLATIVE
that	DET	CENTRAL DEMONSTRATIVE SG
all	DET	PREDETERMINER SG/PL QUANTIFIER
dog's	N	GENITIVE SG
furniture	N	NOMINATIVE SG NOINDEFDETERMINER
one-third	NUM	SG
she	PRON	PERSONAL FEMININE NOMINATIVE SG3
show	V	IMPERATIVE VFIN
show	V	PRESENT -SG3 VFIN
show	N	NOMINATIVE SG
shown	PCP2	SVOO SVO SV
occurred	PCP2	SV
occurred	V	PAST VFIN SV

ENGTWOL Tagging: Stage 1

- First Stage: Run words through FST morphological analyzer to get POS info from morph
- E.g.: Pavlov had shown that salivation …

Pavlov	**PAVLOV N NOM SG PROPER**
had	**HAVE V PAST VFIN SVO**
	HAVE PCP2 SVO
shown	**SHOW PCP2 SVOO SVO SV**
that	ADV
	PRON DEM SG
	DET CENTRAL DEM SG
	CS
salivation	**N NOM SG**

ENGTWOL Tagging: Stage 2

- Second Stage: Apply NEGATIVE constraints
- E.g., Adverbial that rule
  - Eliminate all readings of that except the one in It isn't *that* odd.

Given input: that

If

(+1 A/ADV/QUANT) **; if next word is adj/adv/quantifier**

(+2 SENT-LIM) **; followed by E-O-S**

(NOT -1 SVOC/A) **; and the previous word is not a verb like consider which allows adjective complements (e.g. I consider that odd)**

Then eliminate non-ADV tags

Else eliminate ADV

Transformation-Based (Brill) Tagging

- Combines Rule-based and Stochastic Tagging
  - Like rule-based because rules are used to specify tags in a certain environment
  - Like stochastic approach because we use a tagged corpus to find the best performing rules
    - *Rules are learned from data*
- Input:
  - Tagged corpus
  - Dictionary (*with most frequent tags*)

Transformation-Based Tagging

- Basic Idea: Strip tags from tagged corpus and try to learn them by rule application
  - For untagged, first initialize with most probable tag for each word
  - Change tags according to best rewrite rule, e.g. *"if word-1 is a determiner and word is a verb then change the tag to noun"*
  - Compare to gold standard
  - Iterate
- Rules created via rule templates, e.g.of the form *if word-1 is an X and word is a Y then change the tag to Z"*
  - Find rule that applies correctly to most tags and apply
  - Iterate on newly tagged corpus until threshold reached
  - Return ordered set of rules
- NB: Rules may make errors that are corrected by later rules

Templates for TBL

The preceding (following) word is tagged **z**.
The word two before (after) is tagged **z**.
One of the two preceding (following) words is tagged **z**.
One of the three preceding (following) words is tagged **z**.
The preceding word is tagged **z** and the following word is tagged **w**.
The preceding (following) word is tagged **z** and the word two before (after) is tagged **w**.

#	Change tags From	To	Condition	Example
1	NN	VB	Previous tag is TO	to/TO race/NN → VB
2	VBP	VB	One of the previous 3 tags is MD	might/MD vanish/VBP → VB
3	NN	VB	One of the previous 2 tags is MD	might/MD not reply/NN → VB
4	VB	NN	One of the previous 2 tags is DT	
5	VBD	VBN	One of the previous 3 tags is VBZ	

Sample TBL Rule Application

- Labels every word with its most-likely tag
  - E.g. *race* occurences in the Brown corpus:
    - *P(NN|race) = .98*
    - *P(VB|race)= .02*
    - *is/VBZ expected/VBN to/TO race/NN tomorrow/NN*
- Then TBL applies the following rule
  - "Change NN to VB when previous tag is TO"

    ... *is/VBZ expected/VBN to/TO race/NN tomorrow/NN*

    becomes

    ... *is/VBZ expected/VBN to/TO race/VB tomorrow/NN*

TBL Tagging Algorithm

- Step 1: Label every word with most likely tag (from dictionary)
- Step 2: Check every possible transformation & select one which most improves tag accuracy (cf Gold)
- Step 3: Re-tag corpus applying this rule, and add rule to end of rule set
- Repeat 2-3 until some stopping criterion is reached, e.g., X% correct with respect to training corpus
- RESULT: Ordered set of transformation rules to use on new data tagged only with most likely POS tags

TBL Issues

- Problem: Could keep applying (new) transformations ad infinitum
- Problem: Rules are learned in ordered sequence
- Problem: Rules may interact
- But: Rules are compact and can be inspected by humans

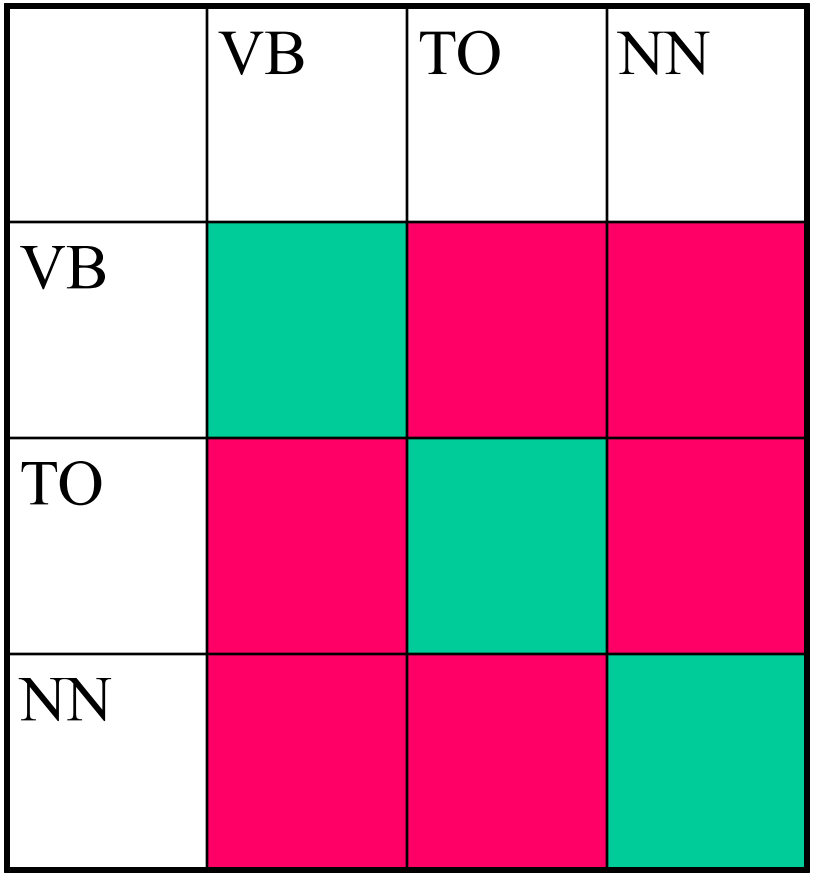
Evaluating Tagging Approaches

- For any NLP problem, we need to know how to evaluate our solutions
- Possible Gold Standards -- ceiling:
  - Annotated naturally occurring corpus
  - Human task performance (96-7%)
    - How well do humans agree?
    - Kappa statistic: avg pairwise agreement corrected for chance agreement
  - Can be hard to obtain for some tasks: sometimes humans don't agree

- Baseline: how well does simple method do?
  - For tagging, most common tag for each word (91%)
  - How much improvement do we get over baseline?

Methodology: Error Analysis

- Confusion matrix:
  - E.g. which tags did we most often confuse with which other tags?
  - How much of the overall error does each confusion account for?



More Complex Issues

- Tag indeterminacy: when ‘truth’ isn’t clear

  Caribbean cooking, child seat
- Tagging multipart words

  wouldn’t --> would/MD n’t/RB
- How to handle unknown words
  - Assume all tags equally likely
  - Assume same tag distribution as all other singletons in corpus
  - Use morphology, word length,….