POS Tagging

Background

- **Part of speech**:
  - Noun, verb, pronoun, preposition, adverb, conjunction, particle, and article
- Recent lists of **POS** (also know as **word classes**, **morphological class**, or **lexical tags**) have much larger numbers of word classes.
  - 45 for Penn Treebank
  - 87 for the Brown corpus, and
  - 146 for the C7 tagset
- The significance of the POS for language processing is that it gives a significant amount of information about the word and its neighbors.
- POS can be used in stemming for IR, since
  - Knowing a word's POS can help tell us which morphological affixes it can take.
  - They can help an IR application by helping select out nouns or other important words from a document.

English Word Classes

- Give a more complete definition of the classes of POS.
  - Traditionally, the definition of POS has been based on morphological and syntactic function.
  - While, it has tendencies toward semantic coherence (e.g., nouns describe people, places, or things and adjectives describe properties), this is not necessarily the case.
- Two broad subcategories of POS:
  1. **Closed class**
  2. **Open class**

English Word Classes

1. **Closed class**
   - Having relatively fixed membership, e.g., prepositions
   - **Function words:**
     - Grammatical words like *of*, *and*, or *you*, which tend to be very short, occur frequently, and play an important role in grammar.
2. **Open class**
   - Four major open classes occurring in the languages of the world: **nouns**, **verbs**, **adjectives**, and **adverbs.**
     - Many languages have no adjectives, e.g., the native American language Lakhota, and Chinese

English Word Classes
Open Class: Noun

- Nouns are traditionally grouped into **proper nouns** and **common nouns.**
  - **Proper nouns:**
    - *Regina, Colorado,* and *IBM*
    - Not preceded by articles, e.g., *the book is upstairs*, but *Regina is upstairs*.
  - **Common nouns**
    - **Count nouns:**
      - Allow grammatical enumeration, i.e., both singular and plural (*goat/goats*), and can be counted (*one goat/ two goats*)
    - **Mass nouns:**
      - Something is conceptualized as a homogeneous group, *snow*, *salt*, and *communism*.
      - Appear without articles where singular nouns cannot (*Snow is white* but not **Goal is white*)

English Word Classes
Open Class: Verb

- **Verbs**
  - Most of the words referring to actions and processes including main verbs like *draw*, *provide*, *differ*, and *go.*
  - A number of morphological forms: non-3rd-person-sg (*eat*), 3rd-person-sg(*eats*), progressive (*eating*), past participle (*eaten*)
  - A subclass: **auxiliaries** (discussed in closed class)

English Word Classes
Open Class: Adjectives

- **Adjectives**
  - Terms describing properties or qualities
  - Most languages have adjectives for the concepts of color (*white, black*), age (*old, young*), and value (*good, bad*), but
  - There are languages without adjectives, e.g., Chinese.

English Word Classes
Open Class: Adverbs

- **Adverbs**

Unfortunately, John walked home extremely slowly yesterday

- Words viewed as modifying something (often verbs)
  - **Directional (or locative) adverbs:** specify the direction or location of some action, *here, downhill*
  - **Degree adverbs:** specify the extent of some action, process, or property, *extremely, very, somewhat*
  - **Manner adverb:** describe the manner of some action or process, *slowly, slinkily, delicately*
  - **Temporal adverbs:** describe the time that some action or event took place, *yesterday, Monday*

English Word Classes
Closed Classes

- Some important closed classes in English
  - **Prepositions:** on, under, over, near, by, at, from, to, with
  - **Determiners:** a, an, the
  - **Pronouns:** she, who, I, others
  - **Conjunctions:** and, but, or, as, if, when
  - **Auxiliary verbs:** can, may, should, are
  - **Particles:** up, down, on, off, in, out, at, by
  - **Numerals:** one, two, three, first, second, third

English Word Classes
Closed Classes: Prepositions

- **Prepositions** occur before nouns, semantically they are relational
  - Indicating spatial or temporal relations, whether literal (*on it, before then, by the house*) or metaphorical (*on time, with gusto, beside herself*)
  - Other relations as well (written *by)*

of	540,085	through	14,964	worth	1,563	pace	12
in	331,235	after	13,670	toward	1,390	nigh	9
for	142,421	between	13,275	plus	750	re	4
to	125,691	under	9,525	till	686	mid	3
with	124,965	per	6,515	amongst	525	o'er	2
on	109,129	among	5,090	via	351	but	0
at	100,169	within	5,030	amid	222	ere	0
by	77,794	towards	4,700	underneath	164	less	0
from	74,843	above	3,056	versus	113	midst	0
about	38,428	near	2,026	amidst	67	o'	0
than	20,210	off	1,695	sans	20	thru	0
over	18,071	past	1,575	circa	14	vice	0

Preposition (and particles) of English from CELEX

English Word Classes
Closed Classes: Particles

- A **particle** is a word that resembles a preposition or an adverb, and that often combines with a verb to form a larger unit call a **phrasal verb**

  So I *went on* for some days cutting and hewing timber …

  Moral reform is the effort to *throw off* sleep …

aboard	aside	besides	forward(s)	opposite	through
about	astray	between	home	out	throughout
above	away	beyond	in	outside	together
across	back	by	inside	over	under
ahead	before	close	instead	overhead	underneath
alongside	behind	down	near	past	up
apart	below	east, etc.	off	round	within
around	beneath	eastward(s),etc.	on	since	without

English single-word particles from Quirk, et al (1985)

English Word Classes
Closed Classes: Auxiliary Verbs

- **Auxiliary verbs:** mark certain semantic feature of a main verb, including
  - whether an action takes place in the present, past or future (tense),
  - whether it is completed (aspect),
  - whether it is negated (polarity), and
  - whether an action is necessary, possible, suggested, desired, etc. (mood).
  - Including **copula** verb *be*, the two verbs *do* and *have* along with their inflection forms, as well as a class of **modal verbs.**

can	70,930	might	5,580	shouldn't	858
will	69,206	couldn't	4,265	mustn't	332
may	25,802	shall	4,118	'll	175
would	18,448	wouldn't	3,548	needn't	148
should	17,760	won't	3,100	mightn't	68
must	16,520	'd	2,299	oughtn't	44
need	9,955	ought	1,845	mayn't	3
can't	6,375	will	862	dare	??
have	???				

English modal verbs from the CELEX on-line dictionary.

English Word Classes
Closed Classes: Others

- **Interjections:** *oh, ah, hey, man, alas*
- **Negatives:** *no, not*
- **Politeness markers:** *please, thank you*
- **Greetings:** *hello, goodbye*
- Existential **there:** *there are two on the table*

Tagsets for English

- There are a small number of popular tagsets for English, many of which evolved from the 87-tag tagset used for the Brown corpus.
  - Three commonly used
    - **The small 45-tag Penn Treebank tagset**
    - The medium-sized 61 tag C5 tageset used by the Lancaster UCREL project's CLAWS tagger to tag the British National Corpus, and
    - The larger 146-tag C7 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 ”)
PP	Personal pronoun	*I, you, he*	(	Left parenthesis	([, (, {, <)
PP$	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*			

Penn Treebank POS tags

Tagsets for English

The grand jury commented on a number of other topics.

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

There are 70 children there

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

Although preliminary findings were reported more than a year ago, the latest results appear in today's New English Journal of Medicine

Although/IN preliminary/JJ findings/NNS were/VBD reported/VBN more/RBR than/IN a/DT year/NN ago/NN ,/, the/DT latest/JJS results/NNS appear/VBP in/IN today/NN 's/POS New/NNP English/NNP Journal/NNP of/IN Medicine/NNP

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	*$*
NNPS	Proper noun, plural	*Carolinas*	#	Pound sign	*#*
PDT	Predeterminer	*all, both*	"	Left quote	(' or ")
POS	Possessive ending	*'s*	"	Right quote	(' or ")
PP	Personal pronoun	*I, you, he*	(	Left parenthesis	([, (, {, <)
PP$	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*			

Part-of-Speech Tagging

- POS tagging (tagging)
  - The process of assigning a POS or other lexical marker to each word in a corpus.
  - Also applied to punctuation marks
  - Taggers play an increasingly important role in speech recognition, NL parsing and IR

Part-of-Speech Tagging

- The input to a tagging algorithm is a string of words and a specified **tagset** of the kind described previously.

```
VB    DT   NN  .
Book that flight .

VBZ  DT  NN   VB     NN   ?
Does that flight serve dinner ?
```

- Automatically assigning a tag to a word is not trivial
  - For example, *book* is ambiguous: it can be a verb or a noun
  - Similarly, *that* can be a determiner, or a complementizer, or an **adverb**
- The problem of POS-tagging is to resolve the ambiguities, choosing the proper tag for the context.

That:

- as a **determiner** (followed by a noun):*Give me that hammer.*
- as a **demonstrative pronoun** (without a following noun):*Who gave you that?*
- as a **conjunction** (connecting two clauses):*I didn't know that she was married.*
- as a **relative pronoun** (forming the subject, object, or complement of a relative clause):*It's a song that my mother taught me.*
- as an **adverb** (before an adjective or adverb):*Three years? I can't wait that long.*

Part-of-Speech Tagging

- How hard is the tagging problem?

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")

The number of word types in Brown corpus by degree of ambiguity.

- Many of the 40% ambiguous tokens are easy to disambiguate, because
  - Various tags associated with a word are not equally likely, or event.
  - E.g., 'a' can be a determiner or a letter (perhaps as part of an acronym)
    - But the determiner sense is much more likely

Part-of-Speech Tagging

- Many tagging algorithms fall into two classes:
  - **Rule-based** taggers
    - Involve a large database of hand-written disambiguation rule specifying, for example, that *an ambiguous word is a noun rather than a verb if it follows a determiner*.
  - **Stochastic** taggers
    - Resolve tagging ambiguities by using a training corpus to *count the probability of a given word having a given tag in a given context*.
- The **Brill tagger**, called the **transformation-based tagger,** shares features of both tagging architecture.

Rule-Based Part-of-Speech Tagging

- The earliest algorithms for automatically assigning POS were based on a two-stage architecture
  - First, use a dictionary to assign each word a list of potential POS.
  - Second, use large lists of hand-written disambiguation rules to winnow down this list to a single POS for each word
- The **ENGTWOL** tagger (1995) is based on the same two stage architecture, with much more sophisticated lexicon and disambiguation rules than before.
  - Lexicon:
    - 56000 entries
    - A word with multiple POS is counted as separate entries

Rule-Based Part-of-Speech Tagging

- In the first stage of tagger,
  - each word is run through the two-level lexicon transducer and
  - the entries for all possible POS are returned.
- A set of about 1,100 constraints are then applied to the input sentences to rule out incorrect POS.

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

N: Noun
V: Verb
PCP2: Past Participle
ADV: Adverb
PRON: Pronoun
DET: Determiner
CS: Subordinating Conjunction

SVOO: Subject-Verb-Object-Object
NOM: non-genitive
VFIN: finite verb
DEM: demonstrative

Rule-Based Part-of-Speech Tagging

- A simplified version of the constraint:

ADVERBIAL-THAT RULE
Given input: “that”
if
  (+1 A/ADV/QUANT); /* *if next word is adj, adverb, or quantifier* */
  (+2 SENT-LIM); /* *and following which is a sentence boundary,* */
  (NOT -1 SVOC/A); /* *and the previous word is not a verb like* */
  /* *‘consider’ which allows adj as object complements* */
then eliminate non-ADV tags
else eliminate ADV tags

- *It isn’t **that** odd.*
- *I considered **that** odd.*

Hidden Markov Model