NLP

Introduction to NLP

Part of speech tagging

The POS Task

- Example
  - Bahrainis vote in second round of parliamentary election
- Jabberwocky (by Lewis Carroll, 1872)

`Twas brillig, and the slithy toves
Did gyre and gimble in the wabe:
All mimsy were the borogoves,
And the mome raths outgrabe.

Parts of Speech

- Open class:
  - nouns, non-modal verbs, adjectives, adverbs
- Closed class:
  - prepositions, modal verbs, conjunctions, particles, determiners, pronouns

Penn Treebank tagset (1/2)

Tag	Description	Example
CC	coordinating conjunction	and
CD	cardinal number	1
DT	determiner	the
EX	existential there	*there* is
FW	foreign word	d‘oeuvre
IN	preposition/subordinating conjunction	in, of, like
JJ	adjective	green
JJR	adjective, comparative	greener
JJS	adjective, superlative	greenest
LS	list marker	1)
MD	modal	could, will
NN	noun, singular or mass	table
NNS	noun plural	tables
NNP	proper noun, singular	John
NNPS	proper noun, plural	Vikings
PDT	predeterminer	*both* the boys
POS	possessive ending	friend*'s*

Penn Treebank tagset (2/2)

Tag	Description	Example
PRP	personal pronoun	I, he, it
PRP$	possessive pronoun	my, his
RB	adverb	however, usually, naturally, here, good
RBR	adverb, comparative	better
RBS	adverb, superlative	best
RP	particle	give *up*
TO	to	*to* go, *to* him
UH	interjection	uhhuhhuhh
VB	verb, base form	take
VBD	verb, past tense	took
VBG	verb, gerund/present participle	taking
VBN	verb, past participle	taken
VBP	verb, sing. present, non-3d	take
VBZ	verb, 3rd person sing. present	takes
WDT	wh-determiner	which
WP	wh-pronoun	who, what
WP$	possessive wh-pronoun	whose
WRB	wh-abverb	where, when

Some Observations

- Ambiguity
  - count (noun) vs. count (verb)
  - 11% of all types but 40% of all tokens in the Brown corpus are ambiguous.
  - Examples
    - *like* can be tagged as ADP VERB ADJ ADV NOUN
    - *present* can be tagged as ADJ NOUN VERB ADV

Some Observations

- More examples:
  - transport, object, discount, address
  - content
- French pronunciation:
  - est, président, fils
- Three main techniques:
  - rule-based
  - machine learning (e.g., conditional random fields, maximum entropy Markov models)
  - transformation-based
- Useful for parsing, translation, text to speech, word sense disambiguation, etc

Example

- Bethlehem/NNP Steel/NNP Corp./NNP ,/, hammered/VBN by/IN higher/JJR **costs/NNS**
- Bethlehem/NNP Steel/NNP Corp./NNP ,/, hammered/VBN by/IN higher/JJR **costs/VBZ**

Sources of Information

- Bethlehem/NNP Steel/NNP Corp./NNP ,/, hammered/ VBN by/IN higher/JJR **costs/NNS**
- Bethlehem/NNP Steel/NNP Corp./NNP ,/, hammered/ VBN by/IN higher/JJR **costs/VBZ**
- Knowledge about individual words
  - lexical information
  - spelling (-or)
  - capitalization (IBM)
- Knowledge about neighboring words

Evaluation

- Baseline
  - tag each word with its most likely tag
  - tag each OOV word as a noun.
  - around 90%
- Current accuracy
  - around 97% for English
  - compared to 98% human performance

Rule-based POS tagging

- Use dictionary or finite-state transducers to find all possible parts of speech
- Use disambiguation rules
  - e.g., ART+V
- Hundreds of constraints can be designed manually

Example in French

<S>	^	beginning of sentence
La	rf b nms u	article
teneur	nfs nms	noun feminine singular
moyenne	jfs nfs v1s v2s v3s	adjective feminine singular
en	p a b	preposition
uranium	nms	noun masculine singular
des	p r	preposition
rivières	nfp	noun feminine plural
,	x	punctuation
bien_que	cs	subordinating conjunction
délicate	jfs	adjective feminine singular
à	p	preposition
calculer	v	verb

Sample Rules

- **BS3 BI1**
  - A BS3 (3rd person subject personal pronoun) cannot be followed by a BI1 (1st person indirect personal pronoun).
  - In the example: “il nous faut” (= “we need”) – “il” has the tag BS3MS and “nous” has the tags [BD1P BI1P BJ1P BR1P BS1P].
  - The negative constraint “BS3 BI1” rules out “BI1P", and thus leaves only 4 alternatives for the word “nous”.
- **N K**
  - The tag N (noun) cannot be followed by a tag K (interrogative pronoun); an example in the test corpus would be: “... fleuve qui ...” (...river, that...).
  - Since “qui” can be tagged both as an ``E'' (relative pronoun) and a “K” (interrogative pronoun), the “E” will be chosen by the tagger since an interrogative pronoun cannot follow a noun (“N”).
- **R V**
  - A word tagged with R (article) cannot be followed by a word tagged with V (verb): for example “l' appelle” (calls him/her).
  - The word “appelle” can only be a verb, but “l'’” can be either an article or a personal pronoun.
  - Thus, the rule will eliminate the article tag, giving preference to the pronoun.

NLP