Natural Language Processing

Part-of-Speech Tagging

Joakim Nivre

Uppsala University
Department of Linguistics and Philology
joakim.nivre@lingfil.uu.se

Parts of Speech

- Basic grammatical categories used since antiquity:
  1. Noun
  2. Verb
  3. Adjective
  4. Adverb
  5. Preposition
  6. Pronoun
  7. Conjunction
  8. Interjection
- Lots of debate in linguistics about their nature and universality
- Nevertheless very robust and useful for NLP

Part-of-Speech Tagging

- Assign a part-of-speech tag to every word of a sentence

Word	Tag
Holmes	PROPN
put	VERB
the	DET
keys	NOUN
on	ADP
the	DET
table	NOUN
.	PUNCT

Why is PoS tagging useful?

- First step in a vast number of practical tasks
  1. Text-to-speech – how to pronounce “lead” or “insult”?
  2. Parsing – need to know if a word is NOUN or VERB
  3. Information extraction – finding names, relations, etc.
- Used as a backoff model for word tokens (sparse data)

Why is PoS tagging hard?

- Lexical ambiguity:
  1. Prince is expected to race/VERB tomorrow
  2. People wonder about the race/NOUN for outer space
- Unknown words:
  1. The rural Babbitt who bloviates about progress and growth

How is it done?

- Lexical information (the word itself)
  - Known words can be looked up in a lexicon listing possible tags for each word
  - Unknown words can be analyzed with respect to affixes, capitalization, special symbols, etc.
- Contextual information (surrounding words)
  - A language model can rank tags in context
- Many different models and techniques (more later)

Quiz 1

- Consider the following incomplete sentence:

  She sent a letter to the ...

- Which parts of speech are likely to occur next?
  1. ADJ
  2. NOUN
  3. VERB

Tag Sets

- There are many potential distinctions we can draw
- Tag sets range from coarse-grained to fine-grained
  1. Universal Dependencies: 17 tags
  2. Penn Treebank, English: 45 tags
  3. SUC, Swedish: 25 tags + features $\approx$ 150 tags
- Choice of tag set may depend on application

Universal Dependencies (UD)

Open class words		Closed class words		Other	
ADJ	adjective	ADP	preposition/postposition	PUNCT	punctuation
ADV	adverb	AUX	auxiliary verb	SYM	symbol
INTJ	interjection	CONJ	coordinating conjunction	X	unspecified
NOUN	noun	DET	determiner		
PROPN	proper noun	NUM	numeral		
VERB	verb	PART	particle		
		PRON	pronoun		
		SCONJ	subordinating conjunction		

Penn Treebank

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*			

How hard is PoS tagging?

		87-tag Original Brown		45-tag Treebank Brown	
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*)

Evaluation

- Evaluation against a manually annotated gold standard
- Evaluation metrics:
  - Accuracy = percentage of correctly tagged tokens
  - Separate results for ambiguous and/or unknown words
- State of the art:
  - 96–98% for English news text
  - What about Turkish?
  - What about Twitter?

Quiz 2

- Consider the following tagging:

  She/PRON won/VERB the/DET race/VERB

- What accuracy score would you give it?
  1. 100%
  2. 75%
  3. 50%

Natural Language Processing

Tagging Methods

Joakim Nivre

Uppsala University
Department of Linguistics and Philology
joakim.nivre@lingfil.uu.se

Part-of-Speech Tagging

- Task:
  - Assign a part-of-speech tag to every word of a sentence
- Useful tools and techniques:
  - Lexicon mapping words to possible tags
  - Linguistic rules for disambiguation in context
  - Statistical models of tags and words in context
  - Heuristics for handling unknown words
- In this lecture:
  - Transformation-based tagging – a rule-based approach
  - HMM tagging – a statistical approach

Transformation-Based Tagging

- Assign each word its most frequent tag

  Prince is expected to race/NOUN tomorrow
  People wonder about the race/NOUN for outer space
  Prince is expected to run/VERB tomorrow
  People wonder about the run/VERB for charity

Transformation-Based Tagging

- Assign each word its most frequent tag

  Prince is expected to race/NOUN tomorrow
  People wonder about the race/NOUN for outer space
  Prince is expected to run/VERB tomorrow
  People wonder about the run/VERB for charity

- Use a sequence of rules to refine the tagging
  1. NOUN $\rightarrow$ VERB if preceding word is to
  2. VERB $\rightarrow$ NOUN if preceding word is the

Transformation-Based Tagging

- Assign each word its most frequent tag

  Prince is expected to race/NOUN tomorrow
  People wonder about the race/NOUN for outer space
  Prince is expected to run/VERB tomorrow
  People wonder about the run/VERB for charity

- Use a sequence of rules to refine the tagging
  1. NOUN $\rightarrow$ VERB if preceding word is to
  2. VERB $\rightarrow$ NOUN if preceding word is the

  Prince is expected to race/VERB tomorrow
  People wonder about the race/NOUN for outer space
  Prince is expected to run/VERB tomorrow
  People wonder about the run/NOUN for charity

Transformation-Based Tagging

- Learning a set of rules from a tagged corpus:
  1. Define a set of rule templates
  2. Assign every word its most frequent tag
  3. Repeat until no further improvement:
     3.1 Apply every rule to the current tagged corpus by itself
     3.2 Add the best rule R to the sequence of rules
     3.3 Transform the current tagged corpus using R

Transformation-Based Tagging

- Learning a set of rules from a tagged corpus:
  1. Define a set of rule templates
  2. Assign every word its most frequent tag
  3. Repeat until no further improvement:
     - 3.1 Apply every rule to the current tagged corpus by itself
     - 3.2 Add the best rule R to the sequence of rules
     - 3.3 Transform the current tagged corpus using R
- Using the rules to tag a new text:
  1. Assign every word its most frequent tag
  2. For every rule $R_1, \ldots, R_n$ in the learned sequence:
     - 2.1 Transform the current tagged corpus using R_i

Quiz 1

- Consider the following initial taggings of the word light:
  1. ... light/VERB the/DET candle/NOUN ...
  2. ... see/VERB the/DET light/VERB ...
  3. ... carry/VERB the/DET light/VERB suitcase/NOUN ...
- And suppose we apply the following two rules in sequence:
  1. VERB $\rightarrow$ NOUN if preceding word is DET
  2. NOUN $\rightarrow$ ADJ if preceding word is DET and following word is NOUN
- Which of the following statements are true?
  1. All three occurrences are correctly tagged in the end
  2. There is at least one error in the end tagging
  3. Removing the second rule gives one more error in the end
  4. Switching the order of the rules has no impact on the end result

Statistical Tagging

- Basic ideas:
  - Build a statistical model of words and their tags
  - Estimate model parameters from (tagged) corpus data
  - Use the model to assign the most probable tags to words
- Example:
  - Part-of-speech tagging using Hidden Markov Models (HMM)

Hidden Markov Models

- Markov models are probabilistic sequence models used for problems such as:
  1. Speech recognition
  2. Spell checking
  3. Part-of-speech tagging
  4. Named entity recognition
- Given a word sequence $w_1, \ldots, w_n$, we want to find the most probable tag sequence $t_1, \ldots, t_n$:

$$\underset{t_1,\ldots,t_n}{\operatorname{argmax}} P(t_1, \ldots, t_n \mid w_1, \ldots, w_n)$$

Model Construction

- Bayesian inversion:

$$P(t_1, \ldots, t_n \mid w_1, \ldots, w_n) = \frac{P(t_1, \ldots, t_n) P(w_1, \ldots, w_n \mid t_1, \ldots, t_n)}{P(w_1, \ldots, w_n)}$$

- Submodels:
  1. Prior: $P(t_1, \ldots, t_n)$
  2. Likelihood: $P(w_1, \ldots, w_n \mid t_1, \ldots, t_n)$
  3. Marginal: $P(w_1, \ldots, w_n)$ – can be ignored in argmax search

Markov Assumptions

- Context model (prior)

$$P(t_1, \ldots, t_n) = \prod_{i=1}^{n} P(t_i | t_{i-k}, \ldots, t_{i-1})$$

- Lexical model (likelihood)

$$P(w_1, \ldots, w_n \mid t_1, \ldots, t_n) = \prod_{i=1}^{n} P(w_i | t_i)$$

Model Parameters

- Contextual probabilities

$$P(t_i|t_{i-k},\ldots,t_{i-1})$$

- Lexical probabilities

$$P(w_i|t_i)$$

- We can estimate these probabilities from a tagged corpus:

$$\hat{P}_{\mathrm{MLE}}(w_i|t_i) = \frac{f(w_i,t_i)}{f(t_i)} \quad \hat{P}_{\mathrm{MLE}}(t_i|t_{i-k},\ldots,t_{i-1}) = \frac{f(t_{i-k},\ldots,t_{i-1},t_i)}{f(t_{i-k},\ldots,t_{i-1})}$$

Computing Probabilities

- The probability of a tagging:

$$P(t_1, \ldots, t_n, w_1, \ldots, w_n) = \prod_{i=1}^{n} P(t_i | t_{i-k}, \ldots, t_{i-1}) P(w_i | t_i)$$

- Finding the most probable tagging:

$$\underset{t_1, \ldots, t_n}{\operatorname{argmax}} \prod_{i=1}^{n} P(t_i | t_{i-k}, \ldots, t_{i-1}) P(w_i | t_i)$$

- This requires an efficient algorithm (more later)

Example

$P(\text{she}\mid\text{PRON})$	$=$	0.1	$P(\text{PRON}\mid\text{START})$	$=$	0.5
$P(\text{can}\mid\text{AUX})$	$=$	0.2	$P(\text{AUX}\mid\text{PRON})$	$=$	0.2
$P(\text{can}\mid\text{NOUN})$	$=$	0.001	$P(\text{NOUN}\mid\text{PRON})$	$=$	0.001
$P(\text{run}\mid VERB)$	$=$	0.01	$P(\text{VERB}\mid\text{AUX})$	$=$	0.5
$P(\text{run}\mid NOUN)$	$=$	0.001	$P(\text{NOUN}\mid\text{AUX})$	$=$	0.001
			$P(\text{VERB}\mid\text{NOUN})$	$=$	0.2
			$P(\text{NOUN}\mid\text{NOUN})$	$=$	0.1

Example

$$
\begin{aligned}
P(\text{she}|\text{PRON}) &= 0.1 \\
P(\text{can}|\text{AUX}) &= 0.2 \\
P(\text{can}|\text{NOUN}) &= 0.001 \\
P(\text{run}|\mathit{VERB}) &= 0.01 \\
P(\text{run}|\mathit{NOUN}) &= 0.001
\end{aligned}
\qquad
\begin{aligned}
P(\text{PRON}|\text{START}) &= 0.5 \\
P(\text{AUX}|\text{PRON}) &= 0.2 \\
P(\text{NOUN}|\text{PRON}) &= 0.001 \\
P(\text{VERB}|\text{AUX}) &= 0.5 \\
P(\text{NOUN}|\text{AUX}) &= 0.001 \\
P(\text{VERB}|\text{NOUN}) &= 0.2 \\
P(\text{NOUN}|\text{NOUN}) &= 0.1
\end{aligned}
$$

$$P(\text{she/PRON can/AUX run/VERB}) = 0.5 \cdot 0.1 \cdot 0.2 \cdot 0.2 \cdot 0.5 \cdot 0.01 = 0.00001$$

Example

$P(\text{she}\mid\text{PRON}) = 0.1$	$P(\text{PRON}\mid\text{START}) = 0.5$
$P(\text{can}\mid\text{AUX}) = 0.2$	$P(\text{AUX}\mid\text{PRON}) = 0.2$
$P(\text{can}\mid\text{NOUN}) = 0.001$	$P(\text{NOUN}\mid\text{PRON}) = 0.001$
$P(\text{run}\mid VERB) = 0.01$	$P(\text{VERB}\mid\text{AUX}) = 0.5$
$P(\text{run}\mid NOUN) = 0.001$	$P(\text{NOUN}\mid\text{AUX}) = 0.001$
	$P(\text{VERB}\mid\text{NOUN}) = 0.2$
	$P(\text{NOUN}\mid\text{NOUN}) = 0.1$

$$P(\text{she/PRON can/AUX run/VERB}) = 0.5 \cdot 0.1 \cdot 0.2 \cdot 0.2 \cdot 0.5 \cdot 0.01 = 0.00001$$

$$P(\text{she/PRON can/NOUN run/NOUN}) = 0.5 \cdot 0.1 \cdot 0.001 \cdot 0.001 \cdot 0.1 \cdot 0.001 = 5 \cdot 10^{-11}$$

Fundamental Problems

- Decoding:
  - How do we compute the best tag sequence given parameters?
- Learning:
  - How do we estimate the parameters?

Quiz 2

- Consider this simple HMM for tagging:

$P(\text{she}\mid\text{PRON}) = 0.1$	$P(\text{PRON}\mid\text{START}) = 0.5$
$P(\text{can}\mid\text{AUX}) = 0.2$	$P(\text{AUX}\mid\text{PRON}) = 0.2$
$P(\text{can}\mid\text{NOUN}) = 0.001$	$P(\text{NOUN}\mid\text{PRON}) = 0.001$
$P(\text{run}\mid VERB) = 0.01$	$P(\text{VERB}\mid\text{AUX}) = 0.5$
$P(\text{run}\mid NOUN) = 0.001$	$P(\text{NOUN}\mid\text{AUX}) = 0.001$
	$P(\text{VERB}\mid\text{NOUN}) = 0.2$
	$P(\text{NOUN}\mid\text{NOUN}) = 0.1$

- Which of the following statements are true?
  1. The probability that can is a NOUN is 0.001.
  2. The probability that the word after an AUX is not a VERB is 0.5.
  3. $P(\text{she/PRON can/AUX}) > P(\text{she/PRON can/NOUN})$