Part-of-Speech Tagging

course based on [Jurafsky and Martin, 2009, Chap.5]

UPPSALA
UNIVERSITET
Marie Dubremetz
marie.dubremetz@lingfil.uu.se

Uppsala, 2016

Presentation Plan

1. **What is PoS Tagging?**

2. **An Example of a Tagged Corpus: SUC**

3. **Evaluation**

4. **Types of Tagging Methods**
    - Rule-based methods
    - Statistical methods

What is a Part-of-speech (PoS)?

Part of Speech: Category of words corresponding to similar grammatical properties.

- traditional parts of speech
  - Noun, verb, adjective, adverb, preposition, article, interjection, pronoun, conjunction, ...
- Variously called:
  - Parts of speech, lexical categories, word classes, morphological classes, lexical tags, ...
- Lots of debate within linguistics about the number, nature, and universality of these
  - We'll completely ignore this debate

PoS 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*
DET	determiner	*the, a, that, those*
INT	interjection	*ouch, hey*
PRO	pronoun	*I, me, mine*
CONJ	conjunction	*and, but, for, because*

Table of Contents

1. **What is PoS Tagging?**
2. **An Example of a Tagged Corpus: SUC**
3. **Evaluation**
4. **Types of Tagging Methods**
   - Rule-based methods
   - Statistical methods

What is PoS Tagging?

Part-of-Speech Tagging, definition.

The process of assigning a part-of-speech tag to every word of a sentence/text

WORD	TAG
the	DET
koala	N
put	V
the	DET
keys	N
on	P
the	DET
table	N

Why PoS-Tagging?

- Distinguish heterophones in speech synthesis
- "I did not object to the object." "To present the present. " "The bandage was wound around the wound."
- Parsing
- Need to know if a word is an N or V before you can parse
- Information extraction
- Finding names, relations, etc.
- Machine translation

What is the challenge in PoS Tagging?

Tag ambiguous words

- Solve the lexical ambiguities
- The/DT **wind/NN** was/VB too/ADV strong/ADJ to/PRP **wind/VB** the/DT sail/NN.

Tag unknown words

The/DT rural/JJ **Babbitt/???** who/WP **bloviates/???** about/IN progress/NN and/CC growth/NN

How is PoS-Tagging done?

Two sources of information

- 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

Two Main approaches

- Rule-based systems
- Statistical systems

Tagsets are not universal

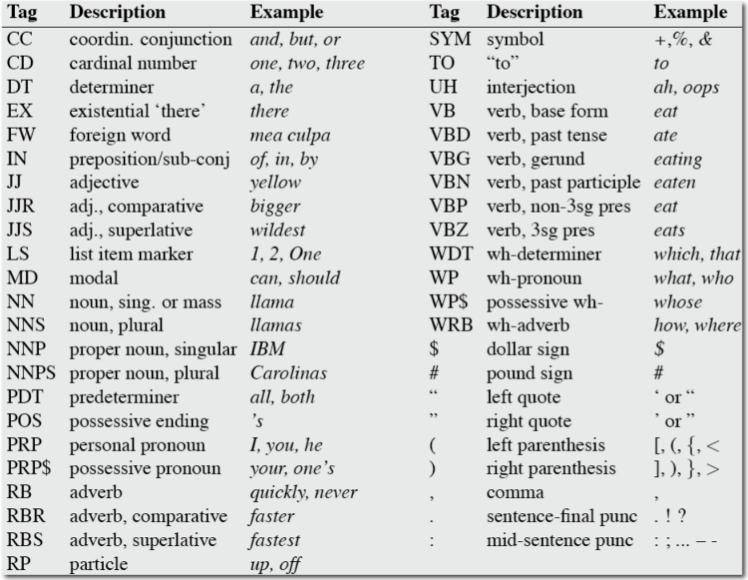
- There are so many potential distinctions we can draw
- To do POS tagging, we need to choose a standard set of tags to work with
- Could pick very coarse tagsets
  - N, V, Adj, Adv, ...
- More commonly used sets are more fine-grained
  - English: Penn Treebank tagset, 45 tags
  - Swedish: SUC tagset, 25 base tags + features $\approx$ 150 tags
- Even more fine-grained tagsets exist

Open and Closed Classes

There are two types of tags.

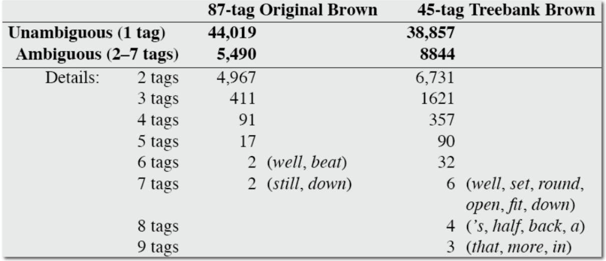
- closed class: a small fixed membership
  - Prepositions: of, in, by, ...
  - Pronouns: I, you, she, mine, his, this, that, ...
  - Determiners: the, a, this, that, ...
  - Usually function words
  - Often frequent and ambiguous
- Open class: new ones can be created all the time
  - English has 4: Nouns, Verbs, Adjectives, Adverbs
  - Usually content words
  - Often rare and (therefore sometimes) unknown

Penn TreeBank PoS 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*			

How Hard is POS Tagging? Measuring Ambiguity



		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*)

Table of Contents

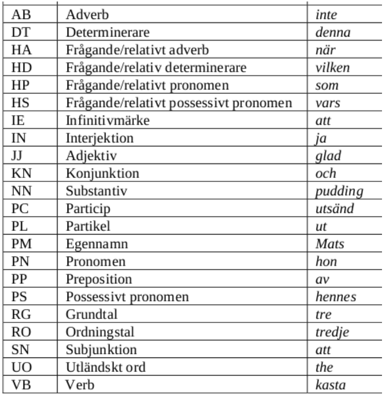
1. What is PoS Tagging?

2. **An Example of a Tagged Corpus: SUC**

3. Evaluation

4. Types of Tagging Methods
   - Rule-based methods
   - Statistical methods

The SUC PoS Tagset



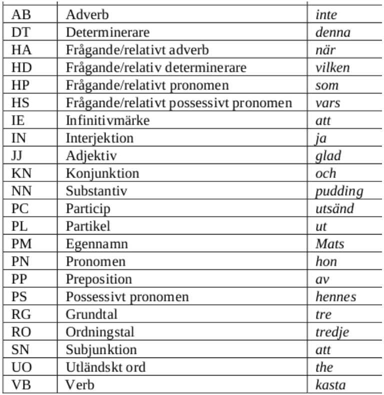
AB	Adverb	*inte*
DT	Determinerare	*denna*
HA	Frågande/relativt adverb	*när*
HD	Frågande/relativ determinerare	*vilken*
HP	Frågande/relativt pronomen	*som*
HS	Frågande/relativt possessivt pronomen	*vars*
IE	Infinitivmärke	*att*
IN	Interjektion	*ja*
JJ	Adjektiv	*glad*
KN	Konjunktion	*och*
NN	Substantiv	*pudding*
PC	Particip	*utsänd*
PL	Partikel	*ut*
PM	Egennamn	*Mats*
PN	Pronomen	*hon*
PP	Preposition	*av*
PS	Possessivt pronomen	*hennes*
RG	Grundtal	*tre*
RO	Ordningstal	*tredje*
SN	Subjunktion	*att*
UO	Utländskt ord	*the*
VB	Verb	*kasta*

QUIZ: Tag me if you can!

Following to the SUC POS Tagset

Tag this:

Och han menade faktiskt allvar



AB	Adverb	*inte*
DT	Determinerare	*denna*
HA	Frågande/relativt adverb	*när*
HD	Frågande/relativ determinerare	*vilken*
HP	Frågande/relativt pronomen	*som*
HS	Frågande/relativt possessivt pronomen	*vars*
IE	Infinitivmärke	*att*
IN	Interjektion	*ja*
JJ	Adjektiv	*glad*
KN	Konjunktion	*och*
NN	Substantiv	*pudding*
PC	Particip	*utsänd*
PL	Partikel	*ut*
PM	Egennamn	*Mats*
PN	Pronomen	*hon*
PP	Preposition	*av*
PS	Possessivt pronomen	*hennes*
RG	Grundtal	*tre*
RO	Ordningstal	*tredje*
SN	Subjunktion	*att*
UO	Utländskt ord	*the*
VB	Verb	*kasta*

QUIZ: Tag me if you can!

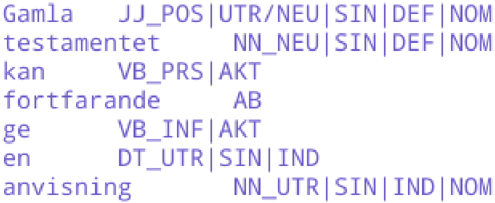
Following to the SUC POS Tagset

Tag this:

Och han menade faktiskt allvar

Och **KN**
han **PN**
menade **VB**
faktiskt **AB**
allvar **NN**

SUC includes morphosyntactic features, as we see in this sample:



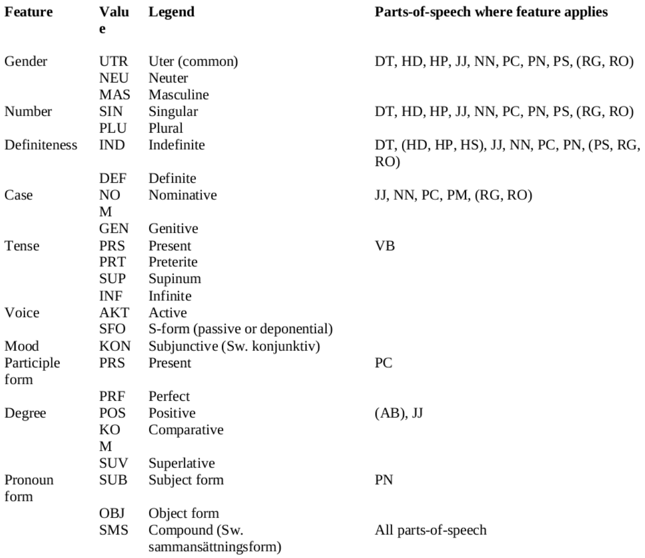
```
Gamla   JJ_POS|UTR/NEU|SIN|DEF|NOM
testamentet     NN_NEU|SIN|DEF|NOM
kan     VB_PRS|AKT
fortfarande     AB
ge      VB_INF|AKT
en      DT_UTR|SIN|IND
anvisning       NN_UTR|SIN|IND|NOM
```

Question

In the next slide you will see the list of morphosyntactic features used in the SUC corpus. Can you add the right morphosyntactic information to the following sample?

Sample:
Och **KN**
han **PN**
menade **VB**
faktiskt **AB**
allvar **NN**

List of the morphosyntactic features



Feature	Value	Legend	Parts-of-speech where feature applies
Gender	UTR	Uter (common)	DT, HD, HP, JJ, NN, PC, PN, PS, (RG, RO)
	NEU	Neuter	
	MAS	Masculine	
Number	SIN	Singular	DT, HD, HP, JJ, NN, PC, PN, PS, (RG, RO)
	PLU	Plural	
Definiteness	IND	Indefinite	DT, (HD, HP, HS), JJ, NN, PC, PN, (PS, RG, RO)
	DEF	Definite	
Case	NOM	Nominative	JJ, NN, PC, PM, (RG, RO)
	GEN	Genitive	
Tense	PRS	Present	VB
	PRT	Preterite	
	SUP	Supinum	
	INF	Infinite	
Voice	AKT	Active	
	SFO	S-form (passive or deponential)	
Mood	KON	Subjunctive (Sw. konjunktiv)	
Participle form	PRS	Present	PC
	PRF	Perfect	
Degree	POS	Positive	(AB), JJ
	KOM	Comparative	
	SUV	Superlative	
Pronoun form	SUB	Subject form	PN
	OBJ	Object form	
	SMS	Compound (Sw. sammansättningsform)	All parts-of-speech

List of the morphosyntactic features

Answer

Och KN
han PN_UTR|SIN|DEF|SUB
menade VB_PRT|AKT
faktiskt AB_POS
allvar NN_NEU|SIN|IND|NOM

Table of Contents

1. What is PoS Tagging?
2. An Example of a Tagged Corpus: SUC
3. **Evaluation**
4. Types of Tagging Methods
   - Rule-based methods
   - Statistical methods

Evaluation

So once you have your PoS tagger running how do you evaluate it?

- Overall error rate with respect to a manually annotated gold-standard test set
- Error rates on known vs. unknown words
- Error rates on particular tags

Accuracy typically reaches 96–97% for English newswire text

Some Vocabulary

Tagging jargon

- **Unknown word**: word that is not in the dictionary/lexicon of the tagger
- **Ambiguous word**: word that can have different tag, depending on the context.
- **Hapax legomenon**: word that appears one time in your corpus.

QUIZ

- You have the following dictionary/lexicon:

  ga *Verb* | *Adv* | *Pronoun*

  bu *Noun*

- You have this corpus:

  ga ga ga bu zo zo mö

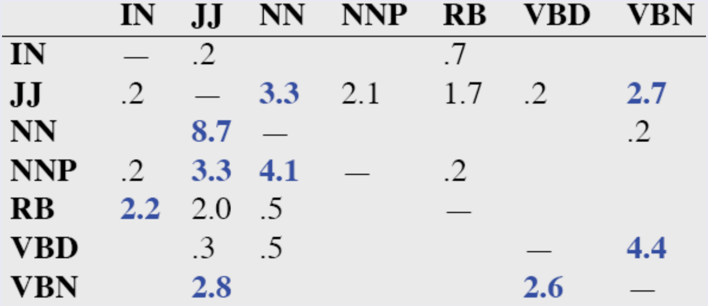
Question

Given this dictionary and this corpus (several ans. possible):

1. The words ‘bu’ and ‘mö’ are hapax legomenon
2. ‘zo’ and ‘mö’ are unknown words
3. ‘ga’ is an ambiguous word
4. all is false

Error Analysis

Look at a confusion matrix



	IN	JJ	NN	NNP	RB	VBD	VBN
IN	—	.2			.7		
JJ	.2	—	**3.3**	2.1	1.7	.2	**2.7**
NN		**8.7**	—				.2
NNP	.2	**3.3**	**4.1**	—	.2		
RB	**2.2**	2.0	.5		—		
VBD		.3	.5			—	**4.4**
VBN		**2.8**				**2.6**	—

See what errors are causing problems

Table of Contents

1. What is PoS Tagging?
2. An Example of a Tagged Corpus: SUC
3. Evaluation
4. **Types of Tagging Methods**
   - Rule-based methods
   - Statistical methods

Two Methods for PoS Tagging

Rule-based systems

- Constraint Grammar
- Transformation-Based Learning

Statistical sequence models

- Hidden Markov Models
- Maximum Entropy Markov Models
- Conditional Random Fields

Table of Contents

1. What is PoS Tagging?

2. An Example of a Tagged Corpus: SUC

3. Evaluation

4. **Types of Tagging Methods**
   - Rule-based methods
   - Statistical methods

Two Methods for PoS Tagging:1)The Rule-Based Systems

Rule-based systems

a) Constraint Grammar
- Assign all possible tags to each word
- Apply rules that discard tags based on context
- Rules created by hand

b) Transformation-Based Learning
- Assign most frequent tag to each word
- Apply rules that replace tags based on context
- Later rules may overwrite earlier rules
- Rules learned from tagged corpus

Two Methods for PoS Tagging:1)The Rule-Based Systems

a) Constraint Grammar

For each ambiguous word, apply a rule. Example: "An ambiguous word is a noun rather than a verb if it succeeds a determiner".

- Advantages:
- Can achieve very high recall with good lexical resources
- Rules can be interpreted by humans, which facilitates debugging
- Drawbacks:
- Not always possible to eliminate all ambiguity
- Rule design is difficult and time-consuming

Two Methods for PoS Tagging:1)The Rule-Based Systems

Here the rules are NOT hand-written and the most probable tags are initially assigned.

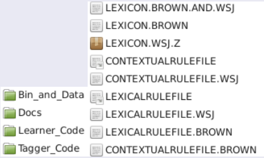
b)Transformation-Based Learning (=Brill tagging)

- Advantages:
- Rules can be interpreted by humans, which facilitates debugging
- Rules are learnt automatically from data
- Drawbacks:
- Not quite as accurate as the best models
- Slow to train on large data sets

QUIZ

This list of file comes from a Tagger.

Can you guess from which kind of tagger those files come? Can you say why?



Bin_and_Data
Docs
Learner_Code
Tagger_Code

LEXICON.BROWN.AND.WSJ
LEXICON.BROWN
LEXICON.WSJ.Z
CONTEXTUALRULEFILE
CONTEXTUALRULEFILE.WSJ
LEXICALRULEFILE
LEXICALRULEFILE.WSJ
LEXICALRULEFILE.BROWN
CONTEXTUALRULEFILE.BROWN

1. A tagger with rules manually written
2. A tagger with rules computationally learned on corpus

Table of Contents

1. What is PoS Tagging?

2. An Example of a Tagged Corpus: SUC

3. Evaluation

4. Types of Tagging Methods
   - Rule-based methods
   - Statistical methods

Two Methods for PoS Tagging:2)Statistical Models

The information is statistics learned from corpus.

We want to answer: What is the most probable tag sequence given a word sequence?

And which is the same as asking:

What is the most probable sequence of tags that generates this sentence?

Exercise: Imagine a corpus tagged by hand.

	zo	ga	bu	bu	zo	bu	zo	zo	
<S>	Adj	V	N	Adj	V	Adj	N	Adj	</S>

What statistical information can you extract from this?

- We can think about extracting the probability of a word to be of a certain PoS-tag (example: $P(Adj|bu)=2/3$).
- It is what the models called ‘discriminative models’ use...
- But it is not the model that we will study in this course
- Think about 2 other types of information to extract.

Exercise: Imagine a corpus tagged by hand.

	zo	ga	bu	bu	zo	bu	zo	zo	
<S>	Adj	V	N	Adj	V	Adj	N	Adj	</S>

What statistical information can you extract from this?

- We can think about extracting the probability of a word to be of a certain PoS-tag (example: $P(Adj|bu)=2/3$).
- It is what the models called ‘discriminative models’ use...
- But it is not the model that we will study in this course
- Think about 2 other types of information to extract.

Exercise: Imagine a corpus tagged by hand.

	zo	ga	bu	bu	zo	bu	zo	zo	
<S>	Adj	V	N	Adj	V	Adj	N	Adj	</S>

What statistical information can you extract from this?

- We can think about extracting the probability of a word to be of a certain PoS-tag (example: $P(Adj|bu)$=2/3).
- It is what the models called 'discriminative models' use...
- But it is not the model that we will study in this course
- Think about 2 other types of information to extract.

Exercise: Imagine a corpus tagged by hand.

	zo	ga	bu	bu	zo	bu	zo	zo	
<S>	Adj	Verb	Noun	Adj	Verb	Adj	Noun	Adj	</S>

1)We can for instance compute this information:
c(Verb,N)=1 c(Noun)=2
$P(Noun|Verb)=1/2$
2)Or this information:
c(Adj,bu)=2 c(Adj)=4
$P(bu|Adj)=2/4$

Exercise: Imagine a corpus tagged by hand.

	zo	ga	bu	bu	zo	bu	zo	zo	
<S>	Adj	Verb	Noun	Adj	Verb	Adj	Noun	Adj	</S>

1)We can for instance compute this information:

$P(Noun|Verb)=1/2$

2)Or this information:

$P(bu|Adj)=2/4$

Instructions

Take the corpus above and:

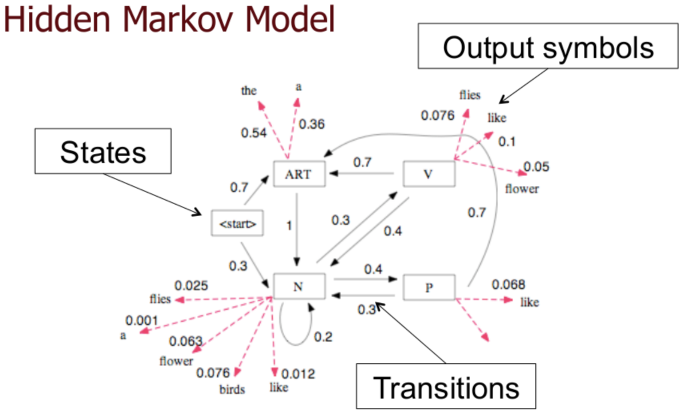
With a red pen, arrow or circle the example of information 1)

With a green pen arrow or circle the example of information 2)

With your own words formulate which kind of information you just captured.

HMM

Congratulations! You just defined the information we need to build a Hidden Markov Model (HMM) for tagging.



Hidden Markov Model

Output symbols

States

Transitions

Hidden Markov Model (HMM): Formally

HMM tagging is based on two mathematical statements

- The Bayesian inference:

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}$$

Applied to tag sequence prediction:

$$P(T_1 ... T_n \mid w_1 ... w_n)$$

$$\Rightarrow \frac{P(T_1 ... T_n) * P(w_1 ... w_n | T_1 ... T_n)}{P(w_1 ... w_n)}$$

- And the Markov assumptions
- Generation of each word w_i, only depends on its tag t_i, and not on previous words
- Generation of each tag t_i only depends on its immediate predecessor t_{i-1}

$$\Rightarrow \prod_{i=1}^{n} P(T_i \mid T_{i-1}) * P(w_i \mid T_i)$$

More Formally

- Alphabet $\Sigma = \{ s_1, s_2, \ldots, s_M \}$

- Set of states $Q = \{ q_1, q_2, \ldots, q_M \}$

- Transition probabilities between any two states

  $a_{ij} = P(q_j \mid q_i)$ = transition prob from state i to state j

- Start probabilities for any state

  $\pi_{0i} = P(q_i)$ = start prob for state I

- Emission probabilities for each symbol and state

  $b_{ik} = P(s_k \mid q_i)$

HMM

We come back on Hidden Markov Model next week!

Summary

Part-of-speech tagging

- Basic step in many analysis pipelines
- Different tagsets for different languages and applications

Methods

- Rule-based systems (Constraint Grammar, Transformation Based Learning)
- Statistical sequence models (HMM, ...)

State of the art

- 96-97% accuracy for English newswire text

To Go Beyond

Don't hesitate to look for key notions (Hidden Markov Model, Constraint Grammar...) in the ACL anthology:
http://aclanthology.info/
This will make you aware of what are those notion used for in our field.

References

Daniel Jurafsky and James H Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, volume 163 of *Prentice Hall Series in Artificial Intelligence*. Prentice Hall, Pearson International Edition, 2009.

Have a look as well here :

https://www.coursera.org/course/nlp