PART OF SPEECH TAGGING

- Introduction
- Part of speech tagging(POS)
- Rule-based taggers
- Statistical taggers
- Hybrid approaches

Content

1. Introduction to Human Language Technology
2. Applications
3. Resources
3. Language models
4. Morphology and lexicons
5. Syntactic processing
6. Semantic processing
7. Generation

INTRODUCTION$_2$

- Parts of speech (POS), word classes, morphological classes, or lexical tags give information about a word and its neighbors

- Since the greeks 8 basic POS have been distinguished:

Noun, verb, pronoun, preposition, adverb, conjunction, adjective, and article

- Modern works use extended lists of POS: 45 in Penn Treebank corpus, 87 in Brown corpus

PART OF SPEECH TAGGING[1]

Tagging is the process of assigning a tag to a word in a corpus

Used for **syntactic processing** and other different tasks:

Speech recognition. Pronunciation may change:

DIS**count* **noun**, *dis**COUNT **verb**

- Information retrieval- morphological affixes
- Lingusitic research- frequency of structures

PART OF SPEECH CATEGORIES$_1$

Closed class. Function words: prepositions, pronouns, determiners,conjunctions, numerals, auxiliary verbs and particles (preposition or adverbs in phrasal verbs)

Open class:

Nouns: people, place and things proper nouns, common nouns,count nouns and mass nouns

Verbs: actions and processes. Main verbs, not auxiliaries

Adjectives: Properties

Adverbs

PART OF SPEECH CATEGORIES$_2$

Brown Corpus tagset (87 tags)

https://en.wikipedia.org/wiki/Brown_Corpus

Penn Treebank tagset (45 tags)

http://web.mit.edu/6.863/www/PennTreebankTa

PART OF SPEECH CATEGORIES$_3$

CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular
NNP	Proper noun, singular
NNS	Noun, plural
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Posessive ending
PRP	Personal pronoun
PP	Possessive pronoun

Penn Tree Bank tagset

RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund
VBN	Verb, past participle
VBP	Verb, non-3rd ps. sing. present
VBZ	Verb, 3rd ps. sing. present
WDT	wh-determiner
WP	wh-pronoun
WP	Possessive wh-pronoun
WRB	wh-adverb

PART OF SPEECH CATEGORIES$_4$

Penn Tree Bank tagset 2

#	Pound sign
$	Dollar sign
.	Sentence-final punctuation
,	Comma
:	Colon, semi-colon
(	Left bracket character
)	Right bracket character
"	Straight double quote
`	Left open single quote
``	Left open double quote
'	Right close single quote
''	Right close double quote

PART OF SPEECH CATEGORIES$_5$

Examples of sentences tagged sentences

Using the 87 tag Brown corpus tagset

Tag **TO** for infinitives

Tag **IN** for prepositional uses of *to*

- *Secretariat/***NNP** *is/***BEZ** *expected/***VBN** ***to***/**TO** race/**VB** *tomorrow/***NR**
- ***to***/**TO** *give/**VB** priority/**NN** **to**/**IN** teacher/**NN** pay/**NN** raises/**NNS***

PART OF SPEECH TAGGING2

PAVLOV **N SG PROPER**

HAVE **V PAST VFIN SVO (verb with subject and object)**

HAVE **PCP2(past participle) SVO**

SHOWN **SHOW PCP2 SVO SV SVOO (verb with subject and two complements)**

THAT **ADV**

PRON DEM SG

DET CENTRAL DEM SG

CS (subordinating conjunction)

SALIVATION **N SG**

PART OF SPEECH TAGGING$_3$

Words taken isolatedly are ambiguous regarding its POS

Yo	bajo	con	el	hombre	bajo	a
PP	VM	SP	TD	NC	VM	NC
	VM				VM	SP
	AQ				AQ	
	NC				NC	
	SP				SP	

tocar	el	bajo	bajo	la	escalera	.
VM	TD	VM	VM	TD	NC	FP
VM		VM	VM	NC		
		AQ	AQ	PP		
		NC	NC			
		SP	SP			

PART OF SPEECH TAGGING 4

Most of words have a unique POS within a context

Yo	bajo	con	el	hombre	bajo	a
PP	**VM**	SP	TD	NC	VM	NC
	VM				VM	SP
	AQ				**AQ**	
	NC				NC	
	SP				SP	

tocar	el	bajo	bajo	la	escalera	.
VM	TD	VM	VM	**TD**	NC	FP
VM		VM	VM	NC		
		AQ	AQ	PP		
		NC	NC			
		SP	**SP**			

PART OF SPEECH TAGGING5

Pos taggers

The goal of a POS tagger is to assign each word the most likely within a context

- Rule-based
- Statistical
- Hybrid

PART OF SPEECH TAGGING$_6$

$W \quad = w_1\, w_2 \ldots w_n$ sequence of words

$T \quad = t_1\, t_2 \ldots t_n$ sequence of POS tags

$$\boxed{f: \; W \;\to\; T \;=\; f(W)}$$

For each word w_i only several of the tags can be assigned (except the unknown words).

We can get them from a lexicon or a morphological analyzer.

Tagset.

Open and closed categories

RULE-BASED TAGGERS$_1$

- Knowledge-driven taggers
- Usually rules built manually
- Limited amount of rules (≈ 1000)
- LM and smoothing explicitely defined.

Brill’s set of templates

“Change tag ***a*** to tag ***b*** when: ..”

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***

a,b,z *and* ***w*** *are part of speech tags*

Rules are **automatically** induced from **tagged corpus**

RULE-BASED TAGGERS$_2$

ADVERBIAL - *THAT* RULE

Given input: "that"

if

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

(+2 SENT-LIM) /* and following is a sentence boundary */

(NOT -1 SVOC/A) /* and the previous word is not a verb like */

/* 'consider' which allows adjs as object complements */

then eliminate non-ADV tags

else eliminate ADV tag

Ex: In the sentence *"I consider that odd "*, that will not be tagged as adverb (ADV)

RULE-BASED TAGGERS$_3$

- + Linguistically motivated rules
- + High precision
  - + ej. EngCG 99.5%
- – High development cost
- – Not transportable
- – Time cost of tagging

• TAGGIT, Green,Rubin,1971
• TOSCA, Oosdijk,1991
• Constraint Grammars, EngCG, Voutilainen,1994, Karlsson et al, 1995
• AMBILIC, de Yzaguirre et al, 2000

RULE-BASED TAGGERS$_4$

Constraint Grammars CG

A CG consists of a sequence of subgrammars each one consisting of a set of restrictions (constraints) which set context conditions

- ej. (@w =0 VFIN (-1 TO))
  - Discards POS VFIN when the previous word is "to"

RULE-BASED TAGGERS$_5$

Constraint Grammars CG

- ENGCG. ENGTWOL

  Reductionist POS tagging

  1,100 constraints

  93-97% of the words are correctly disambiguated

  99.7% accuracy

  Heuristic rules can be applied over the rest 2-3% residual ambiguity with 99.6% precision

  CG syntactic

STATISTICAL POS TAGGING [1]

To find **the most probable tag sequence** given the observation sequence of **n** words $\mathbf{w_1^n}$, that is, find $\mathbf{P(t_1^n|w_1^n)}$ is highest.

But $P(t_1^n|w_1^n)$ is difficult to compute and Bayesian classification rule is used:

$$\mathbf{P(x|y) = P(x)\, P(y|x) / P(y)}$$

When applied to the sequence of words, **the most probable tag sequence** would be

$$\mathbf{P(t_1^n|w_1^n) = P(t_1^n)\, P(w_1^n|t_1^n)/P(w_1^n)}$$

where $P(w_1^n)$ does not change and thus do not need to be calculated

Thus, the **most probable tag sequence** is the product of two probabilites for each possible sequence:

- Prior probability of the tag sequence. Context $\mathbf{P(t_1^n)}$

- Likelihood of the sequence of words considering a sequence of (hidden) tags. $\mathbf{P(w_1^n|t_1^n)}$

STATISTICAL POS TAGGING $_2$

Two simplifications for computing the most probable sequence of tags

- **Prior probability** of the part of speech tag of a word **depends only on** the tag of **the previous** word (bigrams, reduce context to previous). Facilitates the computation of $\mathbf{P(t_1^n)}$

    *Ex. Probability of **noun after determiner***

- **Probability of a word depends only on its part-of-speech tag**. (independent of other words in the context). Facilitates the computation of $\mathbf{P(w_1^n|t_1^n)}$, Likelihood probability.

    *Ex. given the tag **noun**, probabilty of word **dog***

This stochastic algorithm is also called HIDDEN MARKOV MODEL

STATISTICAL POS TAGGING $_3$

Computing the most-likely tag sequence:

- Secretariat/NNP is/BEZ expected/VBN to/TO ***race***/***VB*** tomorrow/NR

- People/NNS continue/VB to/TO inquire/VB the/AT reason/NN for/IN the/AT ***race/NN*** for/IN outer/JJ space/NN

STATISTICAL POS TAGGING 4

Hidden Markov Models (HMM) are extensions of finite state automata

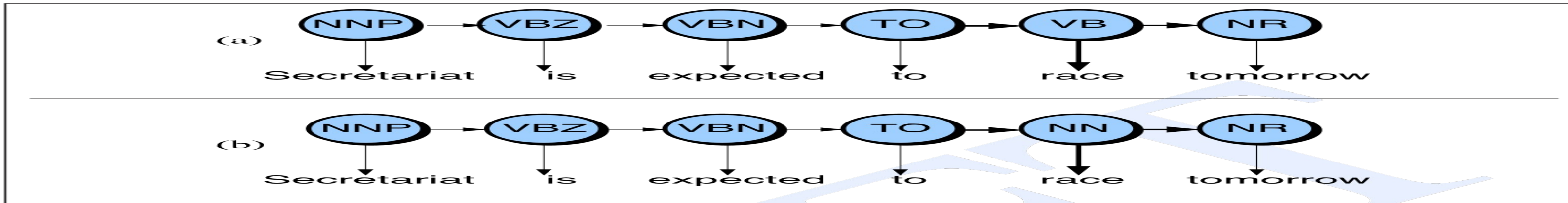

Figure 5.12 Two of the possible sequences of tags corresponding to the Secretariat sentence, one of them corresponding to the correct sequence, in which *race* is a VB. Each arc in these graphs would be associated with a probability. Note that the two graphs differ only in 3 arcs, hence in 3 probabilities.

saw in the previous section, the maximum likelihood estimate for these probabilities can be derived from corpus counts.

Since the (87-tag Brown tagset) tag TO is used only for the infinitive marker *to*, we expect that only a very small number of nouns can follow this marker (as an exercise, try to think of a sentence where a noun can follow the infinitive marker use of *to*). Sure enough, a look at the (87-tag) Brown corpus gives us the following probabilities, showing that verbs are about 500 times as likely as nouns to occur after TO:

$$P(\mathrm{NN}|\mathrm{TO}) = .00047$$
$$P(\mathrm{VB}|\mathrm{TO}) = .83$$

Let's now turn to $P(w_i|t_i)$, the lexical likelihood of the word *race* given a part-of-speech tag. For the two possible tags VB and NN, these correspond to the probabilities $P(\mathrm{race}|\mathrm{VB})$ and $P(\mathrm{race}|\mathrm{NN})$. Here are the lexical likelihoods from Brown:

$$P(\mathrm{race}|\mathrm{NN}) = .00057$$
$$P(\mathrm{race}|\mathrm{VB}) = .00012$$

Finally, we need to represent the tag sequence probability for the following tag (in this case the tag NR for *tomorrow*):

$$P(\mathrm{NR}|\mathrm{VB}) = .0027$$
$$P(\mathrm{NR}|\mathrm{NN}) = .0012$$

Formalization of a Hidden Markov Model

$Q = q_1 q_2 \ldots q_N$ a set of N states

$A = a_{11} a_{12} \ldots a_{n1} \ldots a_{nn}$ a transition probability matrix **A**, each a_{ij} representing the probability of moving from state i to state j, $\sum_{j=1}^{n} a_{ij} = 1 \ \forall i$

$O = o_1 o_2 \ldots o_T$ a sequence of **T** observations, each one drawn from a vocabulary **$V = v_1, v_2, \ldots, v_V$**

$B = b_i(o_t)$ A sequence of observation likelihoods, also called emission probabilities, each expressing the probability of an observation **o_t** being generated from a state **i**.

q0,qF

STATISTICAL POS TAGGING 6

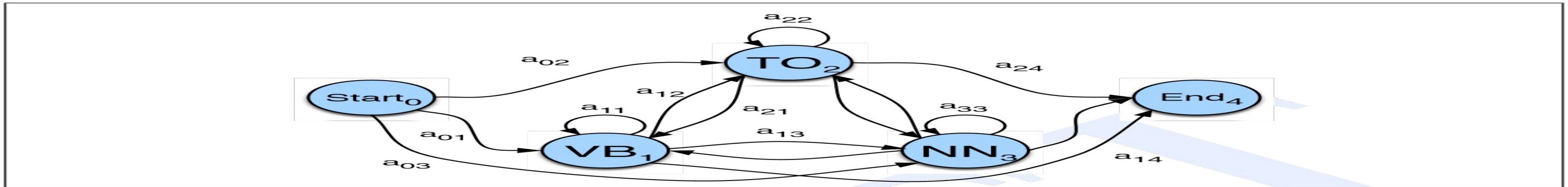

Figure 5.13 The Markov chain corresponding to the hidden states of the HMM. The A transition probabilities are used to compute the prior probability.

An HMM thus has two kinds of probabilities; the A transition probabilities, and the B observation likelihoods, corresponding respectively to the **prior** and **likelihood** probabilities that we saw in equation (5.31). Fig. 5.13 illustrates the prior probabilities in an HMM part-of-speech tagger, showing 3 sample states and some of the A transition probabilities between them. Fig. 5.14 shows another view of an HMM part-of-speech tagger, focusing on the word likelihoods B. Each hidden state is associated with a vector of likelihoods for each observation word.

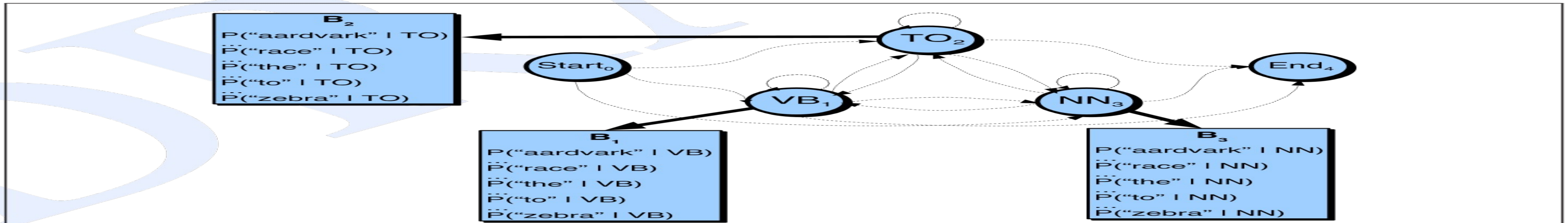

Figure 5.14 The B observation likelihoods for the HMM in the previous figure. Each state (except the non-emitting Start and End states) is associated with a vector of probabilities, one likelihood for each possible observation word.

STATISTICAL POS TAGGING 7

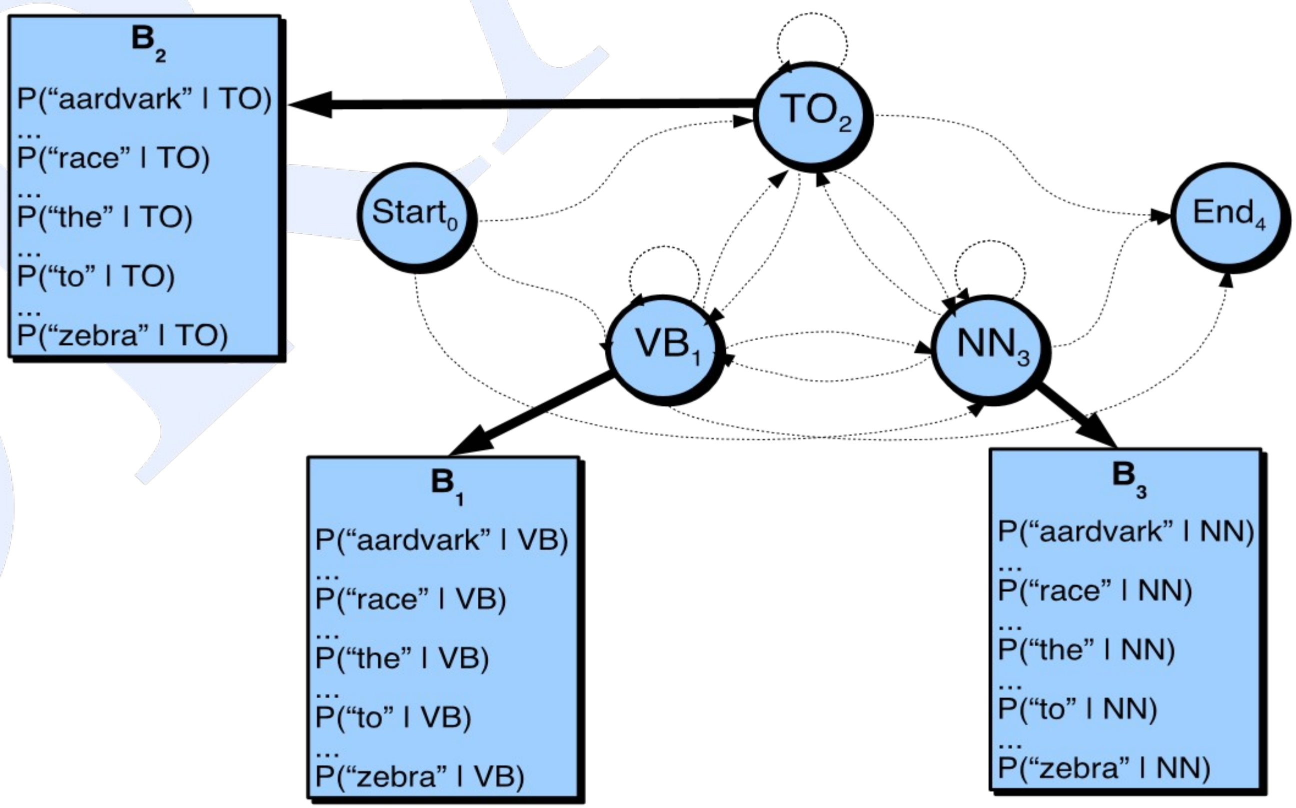

Figure 5.14 The *B* observation likelihoods for the HMM in the previous figure. Each state (except the non-emitting Start and End states) is associated with a vector of probabilities, one likelihood for each possible observation word.

STATISTICAL POS TAGGING 8

	VB	TO	NN	PPSS
<s>	.019	.0043	.041	.067
VB	.0038	.035	.047	.0070
TO	.83	0	.00047	0
NN	.0040	.016	.087	.0045
PPSS	.23	.00079	.0012	.00014

Tag transition probabilities (the **matrix A**, $p(t_i|t_{i-1})$) computed from the 87-tag Brown corpus without smoothing. The rows are labeled with the conditioning event; thus P(PPSS|VB) is .0070. The symbol <s> is the start-of-sentence symbol

STATISTICAL POS TAGGING $_9$

	I	want	to	race
VB	0	.0093	0	.00012
TO	0	0	.99	0
NN	0	.000054	0	.00057
PPSS	.37	0	0	0

Observation likelihoods (the **matrix B**) computed from the 87-tag Brown corpus without smoothing

Hidden Markov Model

Statistical inference. (**Bayesian** inference)

Hidden States associated to n-grams

- Transition probabilities restricted to valid transitions
- Emision probabilities restricted by lexicons

STATISTICAL POS TAGGING 11

- + Well founded theoretical framework
- + Simple models.
- + Acceptable precision
  - + > 97%
- + Language independent
- – Learning the model
  - – Sparseness
- – less precision

- CLAWS, Garside et al, 1987
- De Rose, 1988
- Church, 1988
- Cutting et al,1992
- Merialdo, 1994

STATISTICAL POS TAGGING 12

Data-driven

LM and smoothing automatically learned from tagged corpora (supervised learning)

N-gram

Hidden Markov Models

Machine Learning

Supervised learning

– Semi-supervised

– Forward-Backward, Baum-Welch

Charniak, 1993
Jelinek, 1998
Manning, Schütze, 1999

HYBRID SYSTEMS$_1$

- Transformation-based, error-driven
  - Based on rules automatically acquired

Brill, 1995
Roche,Schabes, 1995

- Maximum Entropy
  - Combination of several knowledge sources
  - No independence is assumed
  - A high number of parameters is allowed (e.g. lexical features)

Ratnaparkhi, 1998,
Rosenfeld, 1994
Ristad, 1997

HYBRID SYSTEMS$_2$

Brill's system

- Based on transformation rules that correct errors produced by an initial HMM tagger
- rule
  - change label A into label B when ...
  - Each rule corresponds to the instantiation of a templete
- templetes
  - The previous (following) word is tagged with Z
  - One of the two previous (following) words is tagged with Z
  - The previous word is tagged with Z and the following with W
  - ...
- Learning of the variables A,B,Z,W through an iterative process That choose at each iteration the rule (the instanciation) correcting more errors.

OTHER COMPLEX SYSTEMS$_1$

- Decision trees
  - Supervised learning
  - ej. TreeTagger

Black,Magerman, 1992
Magerman 1996
Màrquez, 1999
Màrquez, Rodríguez, 1997

Case-based, Memory-based Learning

TiMBL
Daelemans et al, 1996

- Relaxation labelling
  - Statistical and linguistic constraints ej. RELAX

Padrò, 1997

OTHER COMPLEX SYSTEMS$_2$

Combining taggers

- Combination of Language models in a tagger
  - STT+
  - RELAX

Màrquez, Rodríguez, 1998
Màrquez, 1999
Padrò, 1997

- Combination of taggers through votation
  - bootstrapping

Màrquez et al, 1998

- Combinación of classifiers
  - bagging (Breiman, 1996)
  - boosting (Freund, Schapire, 1996)

Brill, Wu, 1998
Màrquez et al, 1999
Abney et al, 1999