Sequence Labeling for Parts of Speech and Named Entities

Erhard Hinrichs

Seminar für Sprachwissenschaft
Eberhard-Karls Universität Tübingen

Three Different Tagsets for American English

- Universal Dependency (UD) Tagset: 17 distinct tags
- Penn Treebank Tagset: 45 distinct tags
- Brown Corpus Tagset: 82 distinct tags

The Stuttgart-Tübingen Tagset (STTS) for German

- tagset with 54 distinct tags
- Tagging guidelines for STTS:
  `https://www.ims.uni-stuttgart.de/documents/ressourcen/lexika/tagsets/stts-1999.pdf`
- For an overview of the tagset see
  `https://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/germantagsets/#id-cfcbf0a7-0`

UD Tagset: English Word Classes (Nivre et al. 2016a)

	Tag	Description	Example
Open Class	**ADJ**	Adjective: noun modifiers describing properties	*red, young, awesome*
	ADV	Adverb: verb modifiers of time, place, manner	*very, slowly, home, today*
	NOUN	words for persons, places, things, etc.	*algorithm, cat, mango, beauty*
	VERB	words for actions and processes	*draw, provide, go*
	PROPN	Proper noun: name of a person, organization, place, etc..	*red, young, awesome*
	INTJ	Interjection: exclamation, greeting, yes/no response, etc.	*oh, um, yes, hello*

UD Tagset: English Word Classes: Continued

	Tag	Description	Example
Closed Class Words	**ADP**	Adposition (Pre-/Postposition): marks a noun's spacial relation	*in, on, by under*
	AUX	Auxiliary: helping verb marking time, place, manner	*can, may, should, are*
	CCONJ	Coordinating Conjunction: joins two phrases/clauses	*and, or, but*
	DET	Determiner: marks noun phrase properties	*a, an, the, this*
	NUM	Numeral	*one, two* first, second

UD Tagset: English Word Classes: Continued

	Tag	Description	Example
Closed Class Words	**PART**	Particle: a preposition-like form used together with a verb	*up, on, off, in, at, by*
	PRON	Pronoun: a shorthand for referring to an entity or event	*she, who, I, others*
	SCONJ	Subordinating Conjunction: joins a main clause with a subordinate clause such as a sentential complement	*that, which*
Other	**PUNCT**	Punctuation	*; , ()*
	SYM	Symbols like $ or emoji	*$, %*
	X	Other	*asdf, qwfg*

Penn Treebank part-of-speech tags

Tag	Description	Example
CC	coord. conj.	*and, but, or*
CD	cardinal number	*one, two*
DT	determiner	*a, the*
EX	existential 'there'	*there*
FW	foreign word	*mea culp*
IN	preposition/subordin-conj	*of, in, by*
JJ	adjective	*yellow*
JJR	comparative adj	*bigger*
JJS	superlative adj	*wildest*
LS	list item marker	*1, 2, One*
MD	modal	*can, should*
NN	sing or mass noun	*llama*

Penn Treebank part-of-speech tags

Tag	Description	Example
NNP	proper noun, sing.	*IBM*
NNPS	proper noun, plu.	*Carolinas*
NNS	noun, plural	*llamas*
PDT	predeterminer	*all, both*
POS	possessive ending	*'s*
PRP	personal pronoun	*I, you, he*
PRP$	possess. pronoun	*your, one's*
RB	adverb	*quickly*
RBR	comparative adv	*faster*
RBS	superlatv. adv	*fastest*
RP	particle	*up, off*
SYM	symbol	+,%, &

Penn Treebank part-of-speech tags

Tag	Description	Example
TO	"to"	*to*
UH	interjection	*ah, oops*
VB	verb base	*eat*
VBD	verb past tense	*ate*
VBG	verb gerund	*eating*
VBN	verb past participle	*eaten*
VBP	verb non-3sg-pr	*eat*
VBZ	verb 3sg pres	*eats*
WDT	wh-determ.	*which, that*
WP	wh-pronoun	*what, who*
WP$	wh-possess.	*whose*
WRB	wh-adverb	*how, where*

Tagging Guidelines for the Penn Treebank

CC or **DT**:	Either/**DT** child could sing.
	Either/**CC** a boy could sing or/**CC** a girl could dance.
CD or **JJ**:	a 50 3/**JJ** victory (cf. a handy/**JJ** victory)
IN or **RP**	the picture we will look at/**IN** next.
	She told off/**RP** her friends.
	She told her friends off/**RP**.
	because/**IN** of/**IN** her late arrival.
	She stepped off/**IN** the train
	* She stepped the train off/**IN**.
IN or **WDT**	the fact that/**IN** you are here.
	a man that/**WDT** I know
JJ or **NP**:	English/**JJ** cuisine tends to be uninspired.
	The English/**NNS** tend to be uninspired cooks.
	The West**JJ** German/**JJ** mark
	He is a West/**NP** German/**NP**.

Tagging Guidelines for the Penn Treebank

JJ or **RB**:	rapid/**JJ** growth/**NN**
	rapid/**JJ** growing/**VBG** plants
JJ or **VBG**:	The conversation became depressing/**JJ**.
	an appetizing/**JJ** dish
	*A dish that appetizes
	an existing/**VBG** safeguards
	safeguards that exist.
JJ or **VBN**:	He became interested/**JJ**.
	He remains guided/**VBN** by these principles.
	They should be kept well-watered/**JJ**.
	At the time, I was married/**JJ**.
NN or **RB**:	Call me when you get home/**RB**.
	Call me when you are at home/**NN**.

Beatrice Santorini (1991). Part-of-Speech Guidelines for the PTB.
`www.cis.upenn.edu/~bies/manuals/tagguide.pdf`

Brown part-of-Speech tags

Tag	Description	Examples
.	sentence closer	. ; ? !
(	left paren	
)	right paren	
*	*not, n't*	
--	dash	
,	comma	
:	colon	
ABL	pre-qualifier	*quite, rather*
ABN	pre-quantifier	*half, all*
ABX	pre-quantifier	*both*
AP	post-determiner	*many, several, next*
AT	article	*a, the, no*

Brown part-of-Speech tags

BE	*be*	
BED	*were*	
BEDZ	*was*	
BEG	*being*	
BEM	*am*	
BEN	*been*	
BER	*are, art*	
BEZ	*is*	
CC	coordinating conjunction	*and, or*
CD	cardinal numeral	*one, two, 2*
CS	subordinating conjunction	*if, although*
DO	*do*	
DOD	*did*	
DOZ	*does*	

Brown part-of-Speech tags

DT	singular determiner	*this, that*
DTI	singular or plural determiner/quantifier	*some, any*
DTS	plural determiner	*these, those*
DTX	determiner/double conjunction	*either*
EX	existentil *there*	
FW	foreign word (hyphenated before regular tag)	
HL	word occurring in headline (hyphenated after regular tag)	
HV	*have*	
HVD	*had* (past tense)	
HVG	*having*	
HVN	*had* (past participle)	
HVZ	*has*	
IN	preposition	
JJ	adjective	
JJR	comparative adjective	
JJS	semantically superlative adjective	*chief, top*
JJT	morphologically superlative adjective	*biggest*

Brown part-of-Speech tags

MD	modal auxiliary	*can, should, will*
NC	cited word (hyphenated after regular tag)	
NN	singular or mass noun	
NN$	possessive singular noun	
NNS	plural noun	
NNS$	possessive plural noun	
NP	proper noun or part of name phrase	
NP$	possessive proper noun	
NPS	plural proper noun	
NPS$	possessive plural proper noun	
NR	adverbial noun	*home, today, west*
NRS	plural adverbial noun	
OD	ordinal numeral	*first, 2nd*

Brown part-of-Speech tags

PN	nominal pronoun	*everybody, nothing*
PN$	possessive nominal pronoun	
PP$	possessive personal pronoun	*my, our*
PP$$	second (nominal) possessive pronoun	*mine, ours*
PPL	singular reflexive/intensive personal pronoun	*myself*
PPLS	plural reflexive/intensive personal pronoun	*ourselves*
PPO	objective personal pronoun	*me, him, it, them*
PPS	3rd. singular nominative pronoun	*he, she, it, one*
PPSS	other nominative personal pronoun	*I, we, they, you*

Brown part-of-Speech tags

QL	qualifier	*very, fairly*
QLP	post-qualifier	*enough, indeed*
RB	adverb	
RBR	comparative adverb	
RBT	superlative adverb	
RN	nominal adverb	*here, then, indoors*
RP	adverb/particle	*about, off, up*
TL	word occurring in title (hyphenated after	
	regular tag)	
TO	infinitive marker *to*	
UH	interjection, exclamation	

Brown part-of-Speech tags

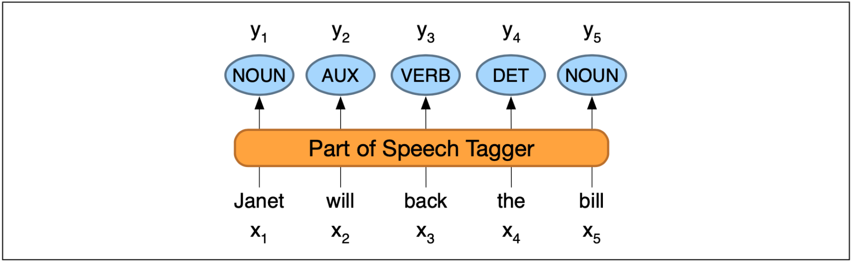
VB	verb, base form	
VBD	verb, past tense	
VBG	verb, present participle/gerund	
VBN	verb, past participle	
VBZ	verb, 3rd. singular present	
WDT	*wh*- determiner	*what, which*
WP$	possessive *wh*- pronoun	*whose*
WPO	objective *wh*- pronoun	*whom, which, that*
WPS	nominative *wh*- pronoun	*who, which, that*
WQL	*wh*- qualifier	*how*
WRB	*wh*- adverb	*how, where, when*

Part-of-Speech Tagging

Example

- There/PRON/EX are/VERB/VBP 70/NUM/CD children/NOUN/NNS there/ADV/RB ./PUNC/.
- Preliminary/ADJ/JJ findings/NOUN/NNS were/AUX/VBD reported/VERB/VBN in/ADP/IN today/NOUN/NN ’s/PART/POS New/PROPN/NNP England/PROPN/NNP Journal/PROPN/NNP of/ADP/IN Medicine/PROPN/NNP

Part-of-Speech Tagging



Tag ambiguity in the Brown and WSJ corpora

Types:	**WSJ**	**Brown**
Unambiguous (1 tag)	44,432 **(86%)**	45,799 **(85%)**
Ambiguous (2+ tags)	7,025 **(14%)**	8,050 **(15%)**
Tokens:		
Unambiguous (1 tag)	577,421 **(45%)**	384,349 **(33%)**
Ambiguous (2+ tags)	711,780 **(55%)**	786,646 **(67%)**

Tag ambiguity in the Brown and WSJ corpora

Example

Citing high fuel prices, [$_{\text{ORG}}$ **United Airlines**] said [$_{\text{TIME}}$ **Friday**] it has increased fares by [$_{\text{MONEY}}$ **$6**] per round trip on flights to some cities also served by lower-cost carriers. [$_{\text{ORG}}$ **American Airlines**], a unit of [$_{\text{ORG}}$ **AMR Corp.**], immediately matched the move, spokesman [$_{\text{PER}}$ **Tim Wagner**] said. [$_{\text{ORG}}$ **United**], a unit of [$_{\text{ORG}}$ **UAL Corp.**], said the increase took effect [$_{\text{TIME}}$ **Thursday**] and applies to most routes where it competes against discount carriers, such as [$_{\text{LOC}}$ **Chicago**] to [$_{\text{LOC}}$ **Dallas**] and [$_{\text{LOC}}$ **Denver**] to [$_{\text{LOC}}$ **San Francisco**].

A list of generic named entity types

Type	**Tag**	**Sample Categories**	**Example sentences**
People	PER	people, characters	**Turing** is a giant of computer science.
Organization	ORG	companies, sports teams	The **IPCC** warned about the cyclone.
Location	LOC	regions, mountains, seas	**Mt. Sanitas** is in **Sunshine Canyon**.
Geo-Political E.	GPE	countries, states	**Palo Alto** is raising the fees for parking.

How difficult is POS tagging in English?

- Roughly 15% of word types are ambiguous.
  - Hence 85% of word types are unambiguous
  - *Janet* is always **PROPN**, *hesitantly* is always **ADV**
- But those 15% tend to be very common.
- So appr. 60% of word tokens are ambiguous.
  - For example:
    back can be an adjective **(JJ)**, a noun (**NN**), a finite (**VBP**) or non-finite verb (**VB**) , an adverb (**RB**), or a particle (**RP**).

Tag ambiguity in the Brown and WSJ corpora

Example

- earnings growth took a **back/JJ** seat
- a small building in the **back/NN**
- a clear majority of senators **back/VBP** the bill
- Dave began to **back/VB** toward the door
- enable the country to buy **back/RP** debt
- I was twenty-one **back/RB** then

Most Frequent Class Baseline

Always compare a classifier against a baseline at least as good as the most frequent class baseline (assigning each token to the class it occurred in most often in the training set).

Examples of type ambiguities in the use of the name Washington

[$_{PER}$ Washington] was born into slavery on the farm of James Burroughs.
[$_{ORG}$ Washington] went up 2 games to 1 in the four-game series.
Blair arrived in [$_{LOC}$ Washington] for what may well be his last state visit.
In June, [$_{GPE}$ Washington] passed a primary seatbelt law.

Named Entities and Named Entity Tagging

Example

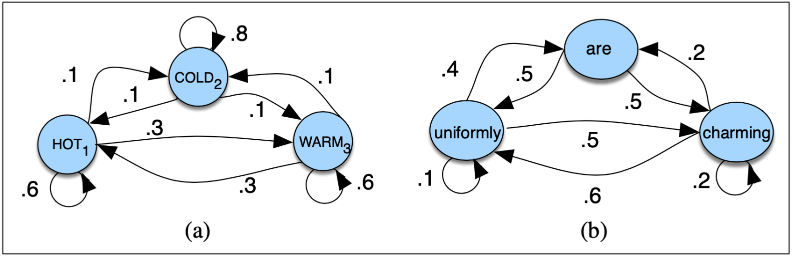
[$_{\text{PER}}$ **Jane Villanueva**] of [$_{\text{ORG}}$ **United**] , a unit of [$_{\text{ORG}}$ **United Airlines Holding**] , said the fare applies to the [$_{\text{LOC}}$ **Chicago**] route.

NER as a sequence model

Words	IO Label	BIO Label	BIOES Label
Jane	I-PER	B-PER	B-PER
Villanueva	I-PER	I-PER	E-PER
of	O	O	O
United	I-ORG	B-ORG	B-ORG
Airlines	I-ORG	I-ORG	I-ORG
Holding	I-ORG	I-ORG	E-ORG
discussed	O	O	O
the	O	O	O
Chicago	I-LOC	B-LOC	S-LOC
route	O	O	O
.	O	O	O

Table: NER as a sequence model, showing IO, BIO, and BIOES taggings

Markov Chains



Markov chain

Formally, a Markov chain is specified by the following components:

$Q = q_1 q_2 \dots q_N$	a set of N **states**
$A = a_{11} a_{12} \dots a_{N1} \dots a_{NN}$	a **transition probability matrix** A, each a_{ij} representing the probability of moving from state i to state j, s.t. $\sum_{j=1}^{n} a_{ij} = 1 \quad \forall i$
$\pi = \pi_1, \pi_2, \dots, \pi_N$	an **initial probability distribution** over states. π_i is the probability that the Markov chain will start in state i. Some states j may have $\pi_j = 0$, meaning that they cannot be initial states. Also, $\sum_{i=1}^{n} \pi_i = 1$

The 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, s.t. $\sum_{j=1}^{n} a_{ij} = 1 \quad \forall i$
$O = o_1 o_2 \ldots o_T$	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 q_i
$\pi = \pi_1, \pi_2, \ldots, \pi_N$	an **initial probability distribution** over states. π_i is the probability that the Markov chain will start in state i. Some states j may have $\pi_j = 0$, meaning that they cannot be initial states. Also, $\sum_{i=1}^{n} \pi_i = 1$

The Hidden Markov Model

A first-order hidden Markov model instantiates two simplifying assumptions. First, as with a first-order Markov chain, the probability of a particular state depends only on the previous state:

$$\textbf{Markov Assumption:} \quad P(q_i|q_1, ..., q_{i-1}) = P(q_i|q_{i-1}) \tag{1}$$

Second, the probability of an output observation o_i depends only on the state that produced the observation q_i and not on any other states or any other observations:

$$\textbf{Output Independence:} \quad P(o_i|q_1, ...q_i, ..., q_T, o_1, ..., o_i, ..., o_T) = P(o_i|q \tag{2}$$

The components of an HMM tagger

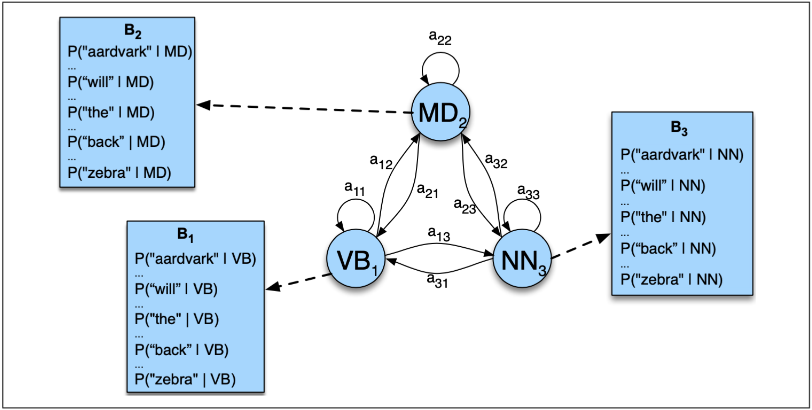
$$P(t_i|t_{i-1}) = \frac{C(t_{i-1}, t_i)}{C(t_{i-1})} \tag{3}$$

$$P(VB|MD) = \frac{C(MD, VB)}{C(MD)} = \frac{10471}{13124} = .80 \tag{4}$$

$$P(w_i|t_i) = \frac{C(t_i, w_i)}{C(t_i)} \tag{5}$$

$$P(will|MD) = \frac{C(MD, will)}{C(MD)} = \frac{4046}{13124} = .31 \tag{6}$$

HMM tagging as decoding



HMM tagging as decoding

$$\widehat{t}_{1:n} = \underset{t_1 \ldots t_n}{\operatorname{argmax}} P(t_1 ... t_n | w_1 ... w_n) \tag{7}$$

The way we'll do this in the HMM is to use Bayes' rule to instead compute:

$$\widehat{t}_{1:n} = \underset{t_1 \ldots t_n}{\operatorname{argmax}} \frac{P(w_1 ... w_n | t_1 ... t_n) P(t_1 ... t_n)}{P(w_1 ... w_n)} \tag{8}$$

$$\widehat{t}_{1:n} = \underset{t_1 \ldots t_n}{\operatorname{argmax}} P(w_1 ... w_n | t_1 ... t_n) P(t_1 ... t_n) \tag{9}$$

$$P(w_1 ... w_n | t_1 ... t_n) \approx \prod_{i=1}^{n} P(w_i | t_i) \quad P(t_1 ... t_n) \approx \prod_{i=1}^{n} P(t_i | t_{i-1}) \tag{10}$$

HMM tagging as decoding

$$\widehat{t}_{1:n} = \underset{t_1 \ldots t_n}{\operatorname{argmax}} P(t_1 ... t_n | w_1 ... w_n) \approx \underset{t_1 \ldots t_n}{\operatorname{argmax}} \prod_{i=1}^{n} \overbrace{P(w_i | t_i)}^{emission} \overbrace{P(t_i | ti - 1)}^{transition} \tag{11}$$

The Viterbi Algorithm

function VITERBI(*observations* of len *T*,*state-graph* of len *N*) **returns** *best-path*, *path-prob*

create a path probability matrix *viterbi[N,T]*
for each state *s* **from** 1 **to** *N* **do** ; initialization step
    $viterbi[s,1] \leftarrow \pi_s * b_s(o_1)$
    $backpointer[s,1] \leftarrow 0$
for each time step *t* **from** 2 **to** *T* **do** ; recursion step
  for each state *s* **from** 1 **to** *N* **do**
    $viterbi[s,t] \leftarrow \max_{s'=1}^{N} \; viterbi[s',t-1] * a_{s',s} * b_s(o_t)$
    $backpointer[s,t] \leftarrow \operatorname{argmax}_{s'=1}^{N} \; viterbi[s',t-1] * a_{s',s} * b_s(o_t)$
$bestpathprob \leftarrow \max_{s=1}^{N} \; viterbi[s,T]$; termination step
$bestpathpointer \leftarrow \operatorname{argmax}_{s=1}^{N} \; viterbi[s,T]$; termination step
bestpath ← the path starting at state *bestpathpointer*, that follows backpointer[] to states back in time
return *bestpath*, *bestpathprob*

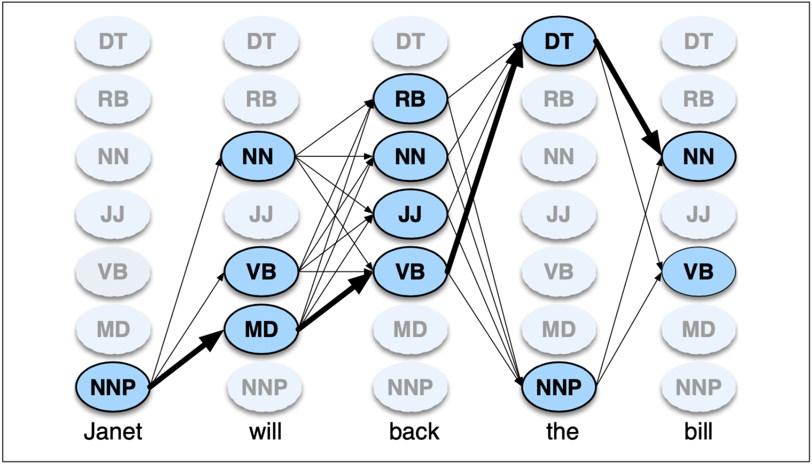
The Viterbi Algorithm

$$v_t(j) = \max_{q_1,\ldots,q_{t-1}} P(q_1 \ldots q_{t-1}, o_1, o_2 \ldots o_t, q_t = j|\lambda) \tag{12}$$

$$v_t(j) = \max_{i=1}^{N} v_{t-1}(i) a_{ij} b_j(o_t) \tag{13}$$

$v_{t-1}(i)$	the **previous Viterbi path probability** from the previous time step
a_{ij}	the **transition probability** y from previous state q_i to current state q_j
$b_j(o_t)$	the **state observation likelihood** of the observation symbol o_t given the the current state j

The Viterbi Algorithm



Working through an example

	NNP	**MD**	**VB**	**JJ**	**NN**	**RB**	**DT**
<s>	0.2767	0.0006	0.0031	0.0453	0.0449	0.0510	0.2026
NNP	0.3777	0.0110	0.0009	0.0084	0.0584	0.0090	0.0025
MD	0.0008	0.0002	0.7968	0.0005	0.0008	0.1698	0.0041
VB	0.0322	0.0005	0.0050	0.0837	0.0615	0.0514	0.2231
JJ	0.0366	0.0004	0.0001	0.0733	0.4509	0.0036	0.0036
NN	0.0096	0.0176	0.0014	0.0086	0.1216	0.0177	0.0068
RB	0.0068	0.0102	0.1011	0.1012	0.0120	0.0728	0.0479
DT	0.1147	0.0021	0.0002	0.2157	0.4744	0.0102	0.0017

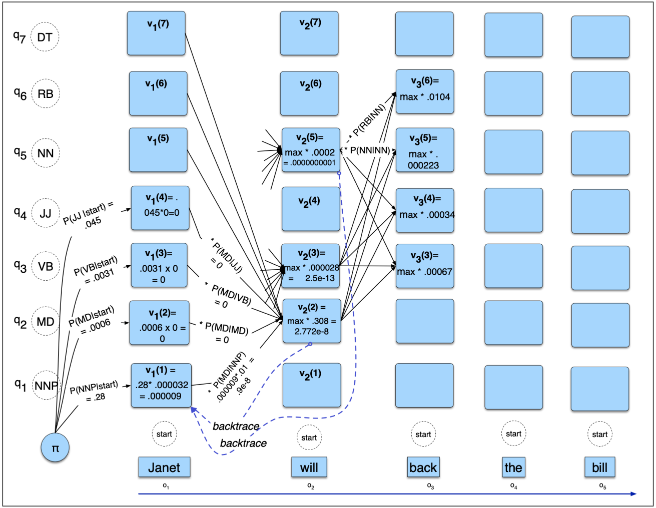
Table: The A transition probabilities $P(t_i|t_{i-1})$ computed from the WSJ corpus without smoothing. Rows are labeled with the conditioning event; thus $P(VB|MD)$ is 0.7968.

Working through an example

	Janet	will	back	the	bill
NNP	0.000032	0	0	0.000048	0
MD	0	0.308431	0	0	0
VB	0	0.000028	0.000672	0	0.000028
JJ	0	0	0.000340	0	0
NN	0	0.000200	0.000223	0	0.002337
RB	0	0	0.010446	0	0
DT	0	0	0	0.506099	0

Table: Observation likelihoods B computed from the WSJ corpus without smoothing, simplified slightly

Working through an example



Conditional Random Fields (CRFs)

$$\begin{aligned}\widehat{Y} &= \underset{Y}{\operatorname{argmax}}\, p(Y|X) \\ &= \underset{Y}{\operatorname{argmax}}\, p(X|Y)p(Y) \\ &= \underset{Y}{\operatorname{argmax}} \prod_i p(x_i|y_i) \prod_i p(y_i|y_{i-1})\end{aligned} \tag{14}$$

CRF to discriminate among the possible tag sequences:

$$\widehat{Y} = \underset{Y \in \mathcal{Y}}{\operatorname{argmax}} P(Y|X) \tag{15}$$

Multinomial logistic regression (Recap from J&M, ch. 5)

- multi-nominal regression, also called *softmax regression* or *maxent classifier*, is used when more than two labels are needed for classification, that is:
  - when we want to label the target variable **y** of each data instance x^i with a unique label **k**, taken from a set of **K** classes.
  - We can represent the correct target value by a **one-hot vector** that assigns the value 1 to the correct class and the value 0 to all other classes.
  - For each prediction $\hat{y}$, we can calculate the probability for each class $\mathbf{k} \in \mathbf{K}$ classes, using the **softmax** function.

$$softmax(z_i) = \frac{exp(z_i)}{\sum_{j=1}^{k} exp(z_j)} \quad 1 \leq i \leq k \tag{16}$$

Conditional Random Fields (CRFs)

Let's assume we have K features, with a weight w_k for each feature F_k:

$$p(Y|X) = \frac{exp\left(\sum_{k=1}^{K} w_k F_k(X, Y)\right)}{\sum_{Y' \in \mathcal{Y}} exp\left(\sum_{k=1}^{K} w_k F_k(X, Y')\right)} \tag{17}$$

Applying Softmax in Logistic Regression

The probability of each output class $\hat{y}_k$ can be computed as:

$$p(\mathbf{y}_k = 1|x) = \frac{exp(\mathbf{w}_k \cdot \mathbf{x} + \mathbf{b}_k)}{\sum_{j=1}^{k} exp(\mathbf{w}_j \cdot \mathbf{x} + b_j)} \tag{18}$$

The vector $\hat{\mathbf{y}}$ of output probabilities for each of the K classes can be computed by:

$$\hat{\mathbf{y}} = softmax(\mathbf{X}\mathbf{w} + \mathbf{b}) \tag{19}$$

Conditional Random Fields (CRFs)

It's common to also describe the same equation by pulling out the denominator into a function $Z(X)$:

$$p(Y|X) = \frac{1}{Z(X)} exp\left(\sum_{k=1}^{K} w_k F_k(X, Y)\right) \tag{20}$$

$$Z(X) = \sum_{Y' \in \mathcal{Y}} exp\left(\sum_{k=1}^{K} w_k F_k(X, Y')\right) \tag{21}$$

$$F_k(X, Y) = \sum_{i=1}^{n} f_k(y_{i-1}, y_i, X, i) \tag{22}$$

Features in a CRF POS Tagger

$$
\begin{aligned}
&\mathbb{1}\{x_i = \textit{the}, y_i = \text{DET}\} \\
&\mathbb{1}\{y_i = \text{PROPN}, x_{i+1} = \textit{Street}, y_{i-1} = \text{NUM}\} \\
&\mathbb{1}\{y_i = \text{VERB}, y_{i-1} = \text{AUX}\}
\end{aligned} \tag{23}
$$

$$
\begin{aligned}
f_{3743} &: y_i = \text{VB and } x_i = \text{back} \\
f_{156} &: y_i = \text{VB and } y_{i-1} = \text{MD} \\
f_{99732} &: y_i = \text{VB and } x_{i-1} = \text{will and } x_{i+2} = \text{bill}
\end{aligned} \tag{24}
$$

Features in a CRF POS Tagger

x_i contains a particular prefix (perhaps from all prefixes of length ≤ 2)
x_i contains a particular suffix (perhaps from all suffixes of length ≤ 2)
x_i's word shape
x_i's short word shape

For example the word *well-dressed* might generate the following non-zero valued feature values:

prefix(x_i) = w
prefix(x_i) = we
suffix(x_i) = ed
suffix(x_i) = d
word-shape(x_i) = xxxx-xxxxxxx
short-word-shape(x_i) = x-x

Features for CRF Named Entity Recognizers

Typical features for a feature-based NER system:

identity of w_i, identity of neighboring words
embeddings for w_i, embeddings for neighboring words
part of speech of w_i, part of speech of neighboring words
presence of w_i in a **gazetteer**
w_i contains a particular prefix (from all prefixes of length ≤ 4)
w_i contains a particular suffix (from all suffixes of length ≤ 4)
word shape of w_i, word shape of neighboring words
short word shape of w_i, short word shape of neighboring words
gazetteer features

Features in a CRF POS Tagger

$\text{prefix}(x_i) = \text{L}$	$\text{suffix}(x_i) = \text{tane}$
$\text{prefix}(x_i) = \text{L'}$	$\text{suffix}(x_i) = \text{ane}$
$\text{prefix}(x_i) = \text{L'O}$	$\text{suffix}(x_i) = \text{ne}$
$\text{prefix}(x_i) = \text{L'Oc}$	$\text{suffix}(x_i) = \text{e}$
$\text{word-shape}(x_i) = \text{X'Xxxxxxxx}$	$\text{short-word-shape}(x_i) = \text{X'Xx}$

Features in a CRF POS Tagger

Words	POS	Short shape	Gazetteer	BIO Label
Jane	NNP	Xx	0	B-PER
Villanueva	NNP	Xx	1	I-PER
of	IN	x	0	O
United	NNP	Xx	0	B-ORG
Airlines	NNP	Xx	0	I-ORG
Holding	NNP	Xx	0	I-ORG
discussed	VBD	x	0	O
the	DT	x	0	O
Chicago	NNP	Xx	1	B-LOC
route	NN	x	0	O
.	.	.	0	O

Table: Some NER features for a sample sentence, assuming that Chicago and Villanueva are listed as locations in a gazetteer. We assume features only take on the values 0 or 1, so the first POS feature, for example, would be represented as $\mathbb{1}\{\text{POS} = \text{NNP}\}$.

Inference and Training for CRFs

$$
\begin{aligned}
\widehat{Y} &= \underset{Y \in \mathcal{Y}}{\operatorname{argmax}} P(Y|X) \\
&= \underset{Y \in \mathcal{Y}}{\operatorname{argmax}} \frac{1}{Z(X)} exp \left(\sum_{k=1}^{K} w_k F_k(X, Y) \right) \\
&= \underset{Y \in \mathcal{Y}}{\operatorname{argmax}} \exp \left(\sum_{k=1}^{K} w_k \sum_{i=1}^{n} f_k(y_{i-1}, y_i, X, i) \right) \\
&= \underset{Y \in \mathcal{Y}}{\operatorname{argmax}} \sum_{k=1}^{K} w_k \sum_{i=1}^{n} f_k(y_{i-1}, y_i, X, i) \\
&= \underset{Y \in \mathcal{Y}}{\operatorname{argmax}} \sum_{i=1}^{n} \sum_{k=1}^{K} w_k f_k(y_{i-1}, y_i, X, i)
\end{aligned}
$$

Inference and Training for CRFs

$$v_t(j) = \max_{i=1}^{N} v_{t-1}(i) a_{ij} b_j(o_t); \quad 1 \leq j \leq N, 1 < t \leq T \tag{25}$$

which is the HMM implementation of

$$v_t(j) = \max_{i=1}^{N} v_{t-1}(i) P(s_j|s_i) P(o_t|s_j) \quad 1 \leq j \leq N, 1 < t \leq T \tag{26}$$

The CRF requires only a slight change to this latter formula, replacing the *a* and *b* prior and likelihood probabilities with the CRF features:

$$v_t(j) = \max_{i=1}^{N} v_{t-1}(i) \sum_{k=1}^{K} w_k f_k(y_{i-1}, y_i, X, i) \quad 1 \leq j \leq N, 1 < t \leq T \tag{27}$$