NATURAL LANGUAGE PROCESSING

UNIT-3

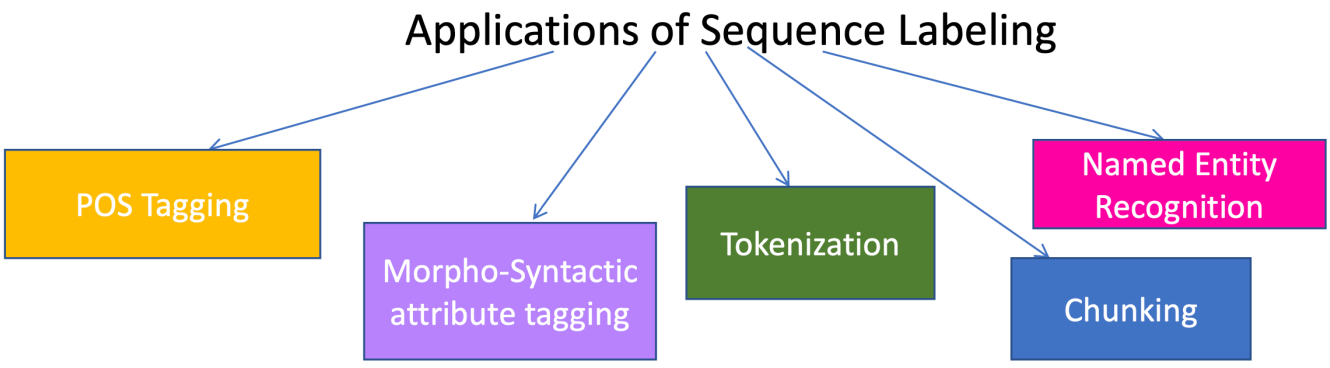
Handling Sequences of Text

feedback/corrections: vibha@pesu.pes.edu

VIBHA MASTI

Sequence Labelling

- Algorithmic assignment of a categorical label to each member of a sequence of observed values
- Set of classification tasks



Generative and Discriminative Classifiers

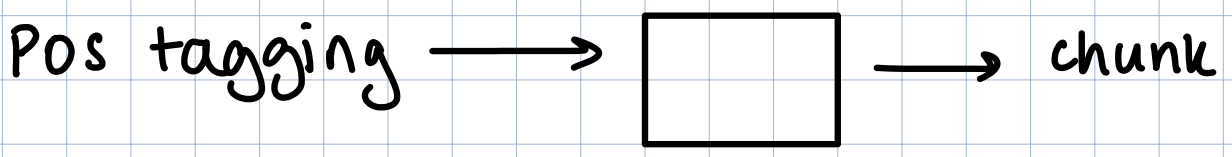
- **Generative classifiers**
- Naïve Bayes
- Bayesian networks
- Markov random fields
- Hidden Markov Models (HMM)

- **Discriminative Classifiers**
- Logistic regression
- Support Vector Machine
- Traditional neural networks
- Nearest neighbour
- Conditional Random Fields (CRF)s
- Maximum Entropy Markov Models (MEMMs)

based on logistic regression & require Viterbi algorithm

chunking

- Extract phrases from unstructured text
- on top of POS tagging



- Standard chunk tags - NP (noun phrase)
  VP (verb phrase)
- POS tags - N (noun)
  V (verb)
  ADV (adverb)

1. POS Tagging

```
Pierre Vinken , 61 years old , will join IBM 's board
as a nonexecutive director Nov. 29 .
```

```
Pierre_NNP Vinken_NNP ,_, 61_CD years_NNS old_JJ ,_,
will_MD join_VB IBM_NNP 's_POS board_NN as_IN a_DT
nonexecutive_JJ director_NN Nov._NNP 29_CD ._.
```

- How word used in sentence
- 8 main POS

1. Noun N
2. Pronoun PRO
3. Adjective ADJ
4. Adverb ADV
5. Verb V
6. Preposition P
7. Conjuction CON
8. Interjection

- Tells us about sentence structure

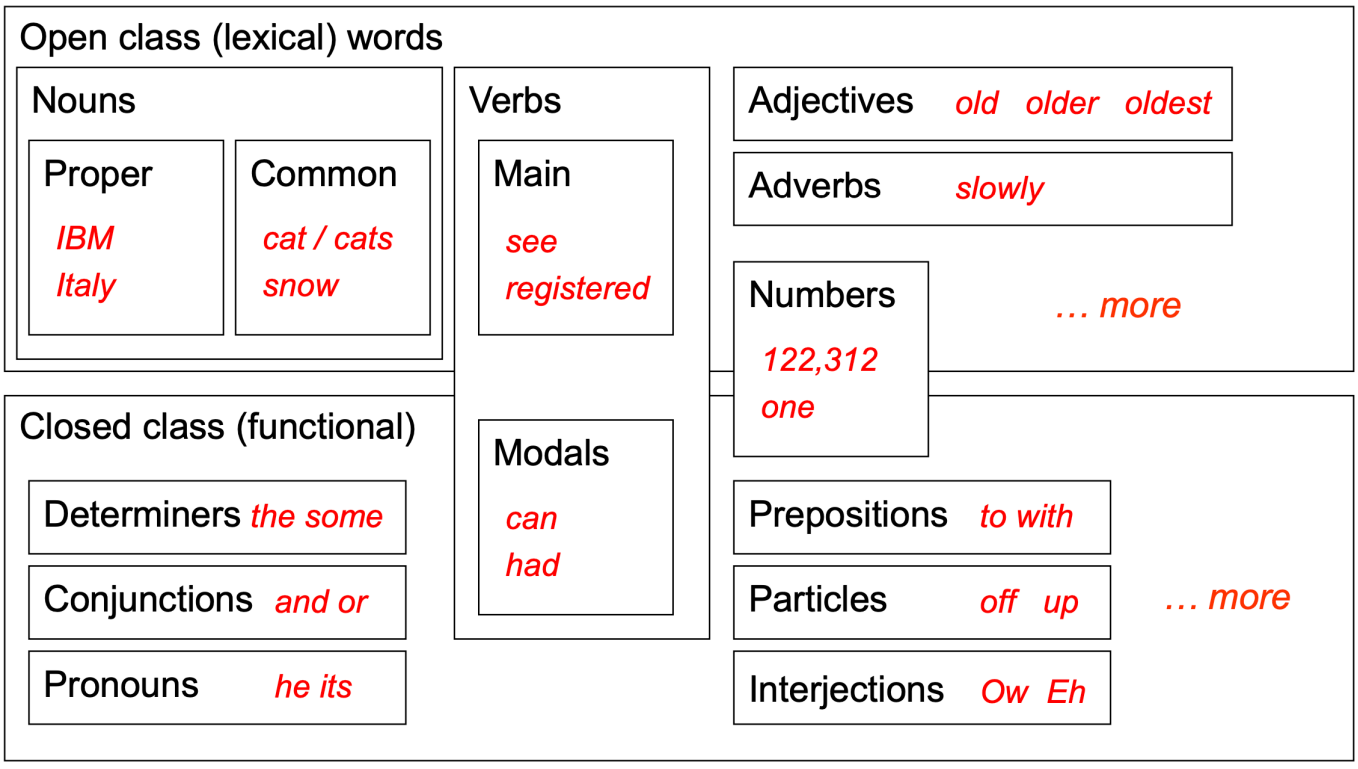
Parts of Speech

1. Closed Class: fixed membership (prepositions, of, it, you etc)
   - Occur frequently

The important closed classes in English include:

- **Prepositions**: on, under, over, near, by, at, from, to, with
- **Particles**: up, down, on, off, in, out, at, by
- **Determiners**: a, an, the, *this/that, these/those, its, our, their*
- **Conjunctions**: and, but, or, as, if, when
- **Pronouns**: she, who, I, others
- **Auxiliary verbs**: can, may, should, is, are, do, have, ...
- **Numerals**: one, two, three, first, second, third, ...

2. Open class: constantly added to (nouns, verbs etc)



Corpora

Year	Name of the Corpus	Size (in words)
1960s - 70s	Brown and LOB	1 Million words
1980s	The Birmingham corpora	20 Million words
1990s	The British National corpus	100 Million words
Early 21st century	The Bank of English corpus	650 Million words

Brown Corpus is a million word collection of samples from 500 written texts from different genres(like newspapers, novels, non-fiction, academics etc.

TreeBank

- Parsed text corpus that annotates syntactic or semantic structure

POS Tags

Number of tags used by different systems/corpora/languages are different

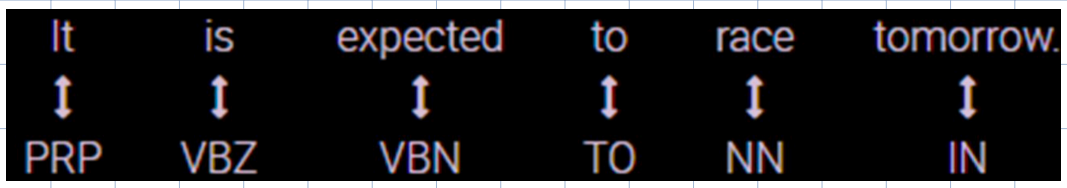
- Penn Treebank (Wall Street Journal Newswire): 45 tags
- Brown corpus (Mixed genres like fiction, biographies, etc): 87 tags
- Lancaster UCREL C5: 61 tags
- Lancaster C7: 145 tags

POS Tagging

- Assign POS marker to each word in an input text

Tag	Description	Example	Tag	Description	Example	Tag	Description	Example
CC	coordinating conjunction	*and, but, or*	PDT	predeterminer	*all, both*	VBP	verb non-3sg present	*eat*
CD	cardinal number	*one, two*	POS	possessive ending	*'s*	VBZ	verb 3sg pres	*eats*
DT	determiner	*a, the*	PRP	personal pronoun	*I, you, he*	WDT	wh-determ.	*which, that*
EX	existential 'there'	*there*	PRP$	possess. pronoun	*your, one's*	WP	wh-pronoun	*what, who*
FW	foreign word	*mea culpa*	RB	adverb	*quickly*	WP$	wh-possess.	*whose*
IN	preposition/ subordin-conj	*of, in, by*	RBR	comparative adverb	*faster*	WRB	wh-adverb	*how, where*
JJ	adjective	*yellow*	RBS	superlatv. adverb	*fastest*	$	dollar sign	$
JJR	comparative adj	*bigger*	RP	particle	*up, off*	#	pound sign	#
JJS	superlative adj	*wildest*	SYM	symbol	+,%, &	"	left quote	' or "
LS	list item marker	*1, 2, One*	TO	"to"	*to*	"	right quote	' or "
MD	modal	*can, should*	UH	interjection	*ah, oops*	(	left paren	[, (, {, <
NN	sing or mass noun	*llama*	VB	verb base form	*eat*	)	right paren	],), }, >
NNS	noun, plural	*llamas*	VBD	verb past tense	*ate*	,	comma	,
NNP	proper noun, sing.	*IBM*	VBG	verb gerund	*eating*	.	sent-end punc	. ! ?
NNPS	proper noun, plu.	*Carolinas*	VBN	verb past part.	*eaten*	:	sent-mid punc	: ; ... – -

Figure 8.1 Penn Treebank part-of-speech tags (including punctuation).



- Disambiguation task
- Eg: book - verb or noun

1) book that flight
2) hand me a book

that - determiner or complementizer

1) does that flight serve dinner
2) I thought that your flight was earlier

"Flies like a flower"

- Flies: noun or verb?
- like: preposition, adverb, conjunction, noun, or verb?
- a: article, noun, or preposition?
- flower: noun or verb?

- **For example the word *back* in following sentences:**

1. The *back* door = JJ(Adjective)
2. On my *back* = NN(Noun singular)
3. Win the voters *back* = RB(Adverb)
4. Promised to *back* the bill = VB(Verb base form)

3 Approaches for POS Tagging

1. Rule-based tagging
   - ENGTWOL
2. Transformation-based tagging
   - Brill
3. Stochastic taggin
   - HMM

1. Rule-Based Tagging

- 2 stage architecture

Stage 1:

- Dictionary of Tagsets
- Assign all possible tags to each word (using morphological / orthographic rules)

Stage 2:

- Write rules to remove selective tags based on preceding, following words
- Leave in correct tag

Q: Sentence = she promised to back the bill

Stage 1:

Dictionary

Word	Tags
she	PRP
promised	VBN, VBD
to	TO
back	VB, JJ, RB, NN
the	DT
bill	NN, VB

personal pronoun

past participle

past tense verb

adverb

Stage 2:

Rule 1: eliminate VBN if VBD is an option when the word follows

"<start> PRP"

			NN		
			RB		
			JJ		VB
PRP	VBD	TO	VB	DT	NN
She	promised	to	back	the	bill

2. Stochastic tagging

- Tag encountered most frequently with the word in training set
- Problem: may yield inadmissible sequence of tags
- Requires training corpus
- Simplest

3. Transformation-Based tagging

- Brill tagging - instance of transformation based learning
- Transformation rules
- Transformation process

  1. If word known, most frequent tag
  2. If word unknown, naively assigns noun

- Change errors with rules

Q: Corpus:

He is expected to race tomorrow.
The race for outer space

1. Using Brown corpus, NN most likely tag for race
2. Use transformation rule to replace NN with VB if race preceded by TO

POS Tagging as a classification problem

- Function $f((w, m), y)$ - feature function for tag type y at position m in sequence $W = (w_1, w_2, \ldots, w_M)$
- Returns a single feature

(w_i, y_i)

w_i: word to be tagged

y_i: tag

Consider the sentence

<s> They can fish </s>

$$f((w = \text{they can fish}, 1), N) = \{(w_m = \text{they}, y_m = N), (w_{m-1} = <s>, y_m = N), (w_{m+1} = \text{can}, y_m = N)\}$$
$$f((w = \text{they can fish}, 2), V) = \{(w_m = \text{can}, y_m = V), (w_{m-1} = \text{they}, y_m = V), (w_{m+1} = \text{fish}, y_m = V)\}$$
$$f((w = \text{they can fish}, 3), V) = \{(w_m = \text{fish}, y_m = V), (w_{m-1} = \text{can}, y_m = V), (w_{m+1} = </s>, y_m = V)\}$$

- Each feature is weighted

  $(w_{m-1} = \text{can}, y_m = V)$: large pos weight

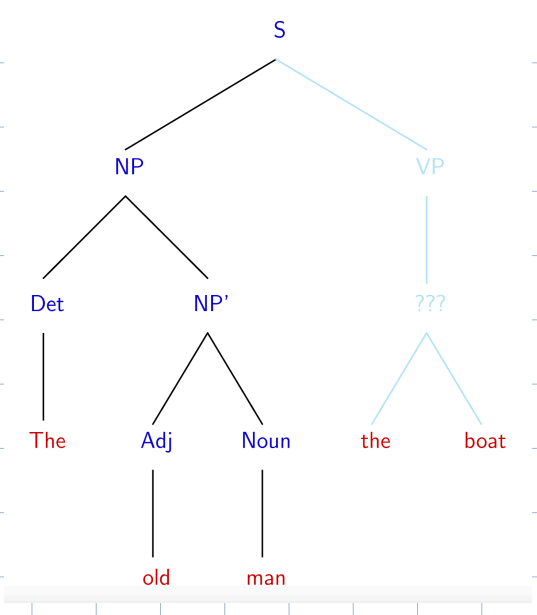
- Verb-verb sequences low preference

Garden Path Sentences

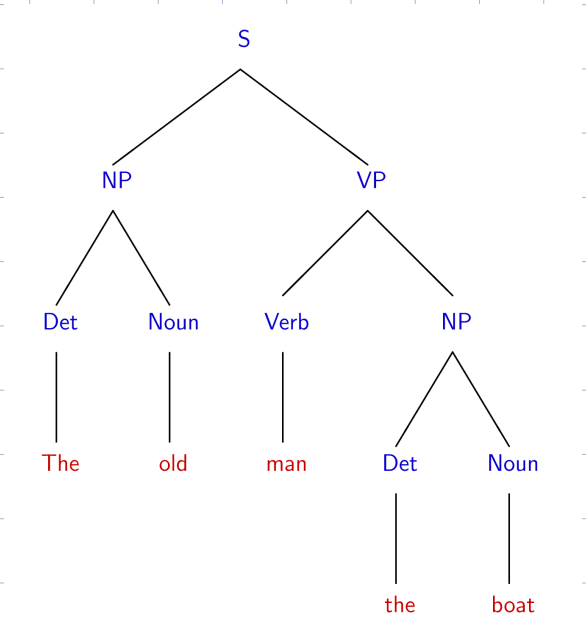
- Grammatically sentence ST reader's most likely interpretation will be incorrect
- Reader's parse yields unintended meaning

Eg 1: "The old man the boat"

- "The old man" expected to be determiner-adjective-noun
- Suggesting uncompletable parse



- Actual (less obvious) parse



- To be interpreted as

The old [people] man the boat.

(operate)

"The complex houses married and single soldiers and their families."

"The horse raced past the barn fell."

2. Named Entity Recognition (NER)

Information Extraction

- Finding factual information in free text
- Facts - structured objects, db records

- *Three bombs have exploded in north-eastern Nigeria, killing 25 people and wounding 12 in an attack carried out by Terrorist group. Authorities said the bombs exploded on Sunday afternoon in the city of Maiduguri*

:Information extracted :

- *TYPE : Crisis ; SUBTYPE : BOMBING ; LOCATION : Maiduguri*
- *DEAD-COUNT : 25 ; INJURED-COINT : 12 ;*
- *PERPETRATOR : Terrorist group; WEAPONS : Bomb*
- *TIME: Sunday afternoon*

- IE tasks
  1. NER
  2. Co-reference resolution
  3. Relation extraction
  4. Event extraction

(a) NER

- ID and classification of types (pre-defined) of named entities
  - Organisations
  - People
  - Places
  - Temporal expression
  - Currency expressions
  - Numerical expressions

- Filling a small-scale template with extracted information

(b) Co-reference resolution

- Multiple co-referring mentions of the same entity in text

Entity mention can be :

- Named, in case an entity is referred to by name; e.g., '*General Electric*' and '*GE*'.
- Pronominal, in case an entity is referred to with a pronoun; e.g., in '*John bought food. But he forgot to buy drinks.*', the pronoun *he* refers to *John*
- Nominal, in case an entity is referred to with a nominal phrase; e.g., in '*Microsoft revealed its earnings. The company also unveiled future plans.*' the definite noun phrase *The company* refers to *Microsoft.*
- Implicit, as in case of using zero-anaphora. Prime minister has visited the place of disaster. [He] flew over with a helicopter

(c) Relation extraction

- Detecting & classifying relationships in text

LocatedIn (Smith, New York): a relation between a person and location, extracted from '*Mr. Smith gave a talk at the conference in New York*'

(d) Event extraction

- Hardest of 4 tasks

Example: The extraction of information on new joint ventures, where the aim is to identify the partners, products, profits and capitalization of the joint venture.

Named Entity

- Anything that has a proper name (person, location, organization)
- Most common tags
  - PER (person),
  - LOC (location),
  - ORG (organization), or
  - GPE (geo-political entity)

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

Problems in Identifying NEs

- Variation of NE (same entity in different form).
  - Manmohan Singh, Manmohan, Dr. Manmohan Singh
- Ambiguity of NE types:
  - 1945 (date vs. time)
  - Washington (location vs. person)
  - May (person vs. month)
  - Tata (person vs. organization)
- Person vs Location
  - Sir C. P Ramaswamy was the Divan of Travancore (Per)
  - Sir C.P Ramaswamy Road is in Chennai (Loc)
- Person vs Organization
  - Anil Ambani opened Reliance Fresh (Per)
  - Reliance Fresh is under Anil Amabani Group Ltd (Org)

Tagset for Named Entity

1. Automatic Content Extraction (ACE)

- Hierarchical tagset

2. CLIA

- Hierarchical tagset
- Developed for Tourism and Health

Named Entity Types

- Hierarchically divided into 3 major entity classes
  1. Name
  2. Time
  3. Numerical expressions
- Examples in slides

Approaches for NER

1. Dictionary / Rule-Based

2. ML
   (a) HMM
   (b) Naive Bayes
   (c) MEMM
   (d) CRFs

3. Hybrid approach

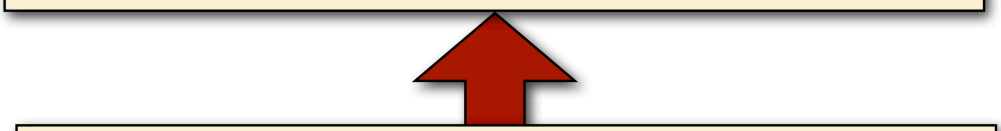
1. Dictionary Approach

- Dicts for identifying NERs
- Gazetteer contains NEs from all domains
  ↑ geographical index or dict
- Tedious to prepare dict
- Simple approach
- Regex for phone, email, captialized names
- Rules for location, context patterns

BIO Encoding

- For noun phrases
- B-NP: beginning of NP chunk
  I-NP: inside NP chunk
  O : outside NP chunk

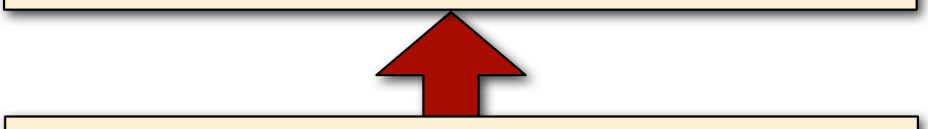
```
[NP Pierre Vinken] , [NP 61 years] old , [VP will join]
[NP IBM] 's [NP board] [PP as] [NP a nonexecutive
director] [NP Nov. 2] .
```



```
Pierre_B-NP Vinken_I-NP ,_O 61_B-NP years_I-NP
old_O ,_O will_B-VP join_I-VP IBM_B-NP 's_O board_B-NP
as_B-PP a_B-NP nonexecutive_I-NP director_I-NP Nov._B-
NP 29_I-NP ._O
```

- Works for any chunk type
- B-PERS: beginning of person chunk
  I-PERS: inside person chunk
  O : outside

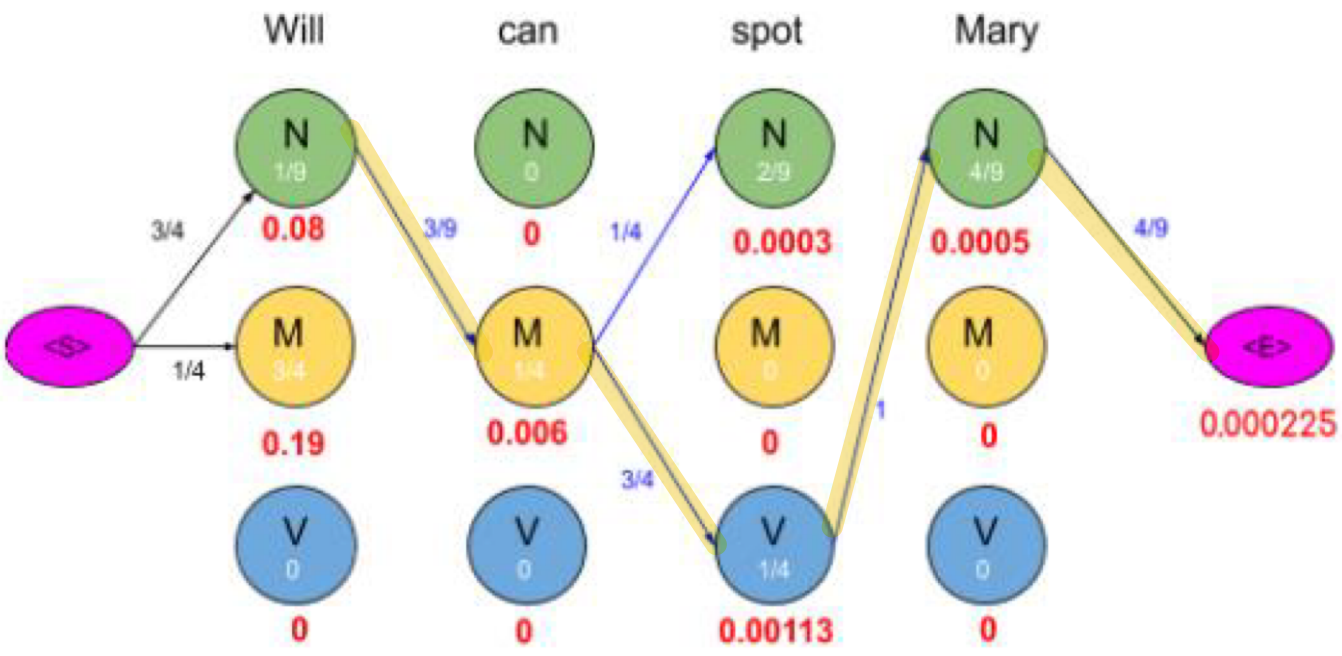
```
[PERS Pierre Vinken] , 61 years old , will join
[ORG IBM] 's board as a nonexecutive director
[DATE Nov. 2] .
```



```
Pierre_B-PERS Vinken_I-PERS ,_O 61_O years_O old_O ,_O
will_O join_O IBM_B-ORG 's_O board_O as_O a_O
nonexecutive_O director_O Nov._B-DATE 29_I-DATE ._O
```

3. HMM

- Decoding problem - most likely sequence of hidden states
- Trellis representation
- Viterbi algorithm – max over prev prob
- Forward algorithm – sum over prev prob



- See MI unit 3

4. Maximum Entropy Models

- Multinomial logistic regression
- Probability of class given the observation x

$$P(c|x) = \frac{1}{Z} e^{\sum_i w_i f_i}$$

class ↗

w_i : weight for feature f_i

- Every observation has a feature vector f
- Features are indicator (binary) variables in vector form
- Eg features:
  - This word ends in -ing
  - Previous word is the
- Each feature f_i is associated with a weight w_i
- Use features to predict part of speech tag
- Let weight of a feature depend on a particular class

$$p(c|x) = \frac{\exp\left(\sum_{i=0}^{N} w_{ci} f_i\right)}{\sum_{c' \in C} \exp\left(\sum_{i=0}^{N} w_{c'i} f_i\right)}$$

Eg: Secretariat/NNP is/VBZ expected/VBN to/TO race/? tomorrow

race: VB or NN

In sequence x:

$\overrightarrow{\text{word}}$: words

$\vec{t}$: tags

Possible features:

$$f_1(c, x) = \begin{cases} 1 & \text{if } \text{word}_i = \text{race} \ \& \ c = \text{NN} \\ 0 & \text{otherwise} \end{cases}$$

$$f_2(c, x) = \begin{cases} 1 & \text{if } t_{i-1} = \text{TO} \ \& \ c = \text{VB} \\ 0 & \text{otherwise} \end{cases}$$

$$f_3(c, x) = \begin{cases} 1 & \text{if } \text{suffix}(\text{word}_i) = \text{ing} \ \& \ c = \text{VBG} \\ 0 & \text{otherwise} \end{cases}$$

- A feature is simply a property dependent on C and observation x
- Weight $w_i(C,x)$ indicates how strong the feature f_i is for the word i

		f1	f2	f3	f4	f5	f6
VB	f	0	1	0	1	1	0
VB	w		.8		.01	.1	
NN	f	1	0	0	0	0	1
NN	w	.8					-1.3

Figure 6.19 Some sample feature values and weights for tagging the word *race* in (6.81).

current word: race

$$P(NN|x) = \frac{e^{.8}e^{-1.3}}{e^{.8}e^{-1.3} + e^{.8}e^{.01}e^{.1}} = .20$$

$$P(VB|x) = \frac{e^{.8}e^{.01}e^{.1}}{e^{.8}e^{-1.3} + e^{.8}e^{.01}e^{.1}} = .80$$

- MEM: probability of a class on a word (not for a whole sequence)
- Turn into a Markov model

5. Maximum Entropy Markov Model

- Previous word tag classification used as feature for next word
- $W = w_1^n$ (sequence of words)

  $T = t_1^n$ (sequence of tags)
- HMM: maximize $P(T|W)$ using Bayes' rule (compute likelihood)

$$\hat{T} = \arg\max_T P(T|W)$$

$$= \arg\max_T P(W|T)\,P(T)$$

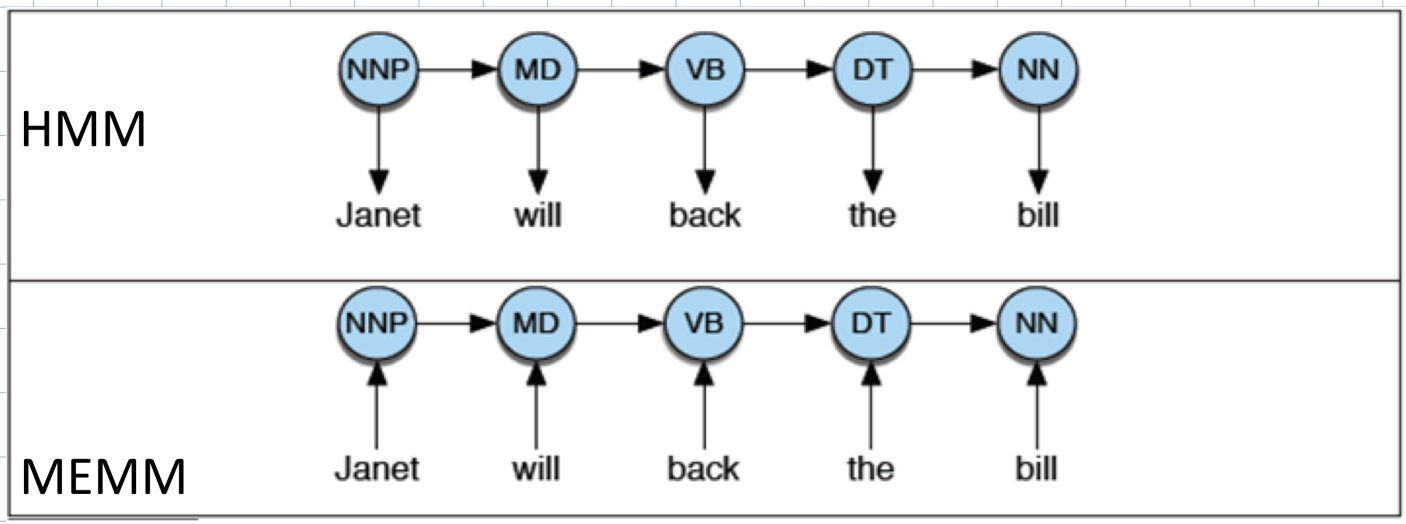
$$= \arg\max_T \prod_i P(\text{word}_i|\text{tag}_i) \prod_i P(\text{tag}_i|\text{tag}_{i-1})$$

emission — transition

- MEMM - use training (directly compute posterior)

$$\hat{T} = \arg\max_T P(T|W)$$

$$= \arg\max_T \prod_i P(t_i|w_i, t_{i-1})$$



- In addition, add features (depending on other previous / next words)
- MEMM POS tagger conditions on
  1. Observation word
  2. Neighbouring words
  3. Previous tags
  4. Combinations
- Feature templates

$\langle t_i, w_{i-2}\rangle, \langle t_i, w_{i-1}\rangle, \langle t_i, w_i\rangle, \langle t_i, w_{i+1}\rangle, \langle t_i, w_{i+2}\rangle, \langle t_i, t_{i-1}\rangle, \langle t_i, t_{i-2}, t_{i-1}\rangle,$

$\langle t_i, t_{i-1} w_i\rangle, \langle t_i, w_{i-1}, w_i\rangle, \langle t_i, w_i, w_{i+1}\rangle$

- Used to automatically populate set of features from every instance
- Features for unknown words

Word Shape Features

- Lowercase letters — x
- Uppercase letters — X
- Digits — d
- Retain punctuation

- Eg: I.M.F ⟶ X.X.X
- Eg: DC10-30 ⟶ XXdd-dd
- Shortened version: remove consecutive char types
- Eg: DC10-30 ⟶ Xd-d

Features for Known Words

- Computed for all training words

Features for Unknown Words

- Can be computed for all training words
- Only on training words with freq < threshold
- Feature cutoff

Decoding & Training MEMMs

$$\hat{T} = \underset{T}{\operatorname{argmax}} \; P(T \mid W)$$

l: neighbourhood words

k: neigh tags

$$= \underset{T}{\operatorname{argmax}} \prod_i P\left(t_i \mid w_{i-l}^{i+l}, t_{i-k}^{i+k}\right)$$

$$= \underset{T}{\operatorname{argmax}} \prod_i \frac{\exp\left(\sum_j \theta_j f_j\left(t_i, w_{i-l}^{i+l}, t_{i-k}^{i+k}\right)\right)}{\sum_{t' \in \text{tagset}} \exp\left(\sum_j \theta_j f_j\left(t', w_{i-l}^{i+l}, t_{i-k}^{i+k}\right)\right)}$$

softmax

Training

1. Greedy
2. Probability (Viterbi algorithm)

Label Bias Problem

https://awni.github.io/label-bias/