Amirkabir University of Technology
(Tehran Polytechnic)

Natural Language Processing

Lecture 10: Parts of Speech Tagging and Named Entity Recognition

Amirkabir University of Technology
Dr Momtazi

Outline

- **Parts of Speech Tagging**
- Named Entity Recognition
- Sequence Modeling
- PGM-based Model for Sequence Labelling
- Evaluation

Parts of Speech

- 8 Parts of speech are traditionally used to summarize the linguistic knowledge
  - Noun, Verb, Preposition, Adverb, Article, Interjection, Pronoun, Conjunction

- The modified list is currently used
  - Noun, Verb, Auxiliary, Preposition, Adjective, Adverb, Number, Determiner, Interjection, Pronoun, Conjunction, Particle

- Known as:
  - Parts of speech
  - Lexical categories
  - Word classes
  - Morphological classes
  - Lexical tags

POS Examples

Noun	book/books, sugar, Germany, Sony
Verb	eat, wrote
Auxiliary	can, should, have
Adjective	new, newer, newest
Adverb	well, urgently
Numbers	872, two, first
Determiner	the, some
Conjunction	and, or
Pronoun	he, my
Preposition	to, in
Particle	off, up
Interjection	Ow, Eh

Open vs. Closed Classes

- Closed class words
  - Relatively fixed membership
  - Usually **function** words: short, frequent words with grammatical function
    - determiners: ***a, an, the***
    - pronouns: ***she, he, I***
    - prepositions: ***on, under, over, near, by, …***

- Open class words
  - Usually **content** words: Nouns, Verbs, Adjectives, Adverbs
    - Plus interjections: **oh, ouch, uh-huh, yes, hello**
  - New nouns and verbs like *iPhone* or *to fax*

Open vs. Closed Classes

Open class ("content")

Nouns

Proper
Janet
Italy

Common
cat, cats
mango

Verbs

Main
eat
went

Auxiliary
can
had

Adjectives *old green tasty*

Adverbs *slowly yesterday*

Interjections *Ow hello*

Numbers
122,312
one

… more

Closed class ("function")

Determiners *the some*

Conjunctions *and or*

Pronouns *they its*

Prepositions *to with*

Particles *off up*

… more

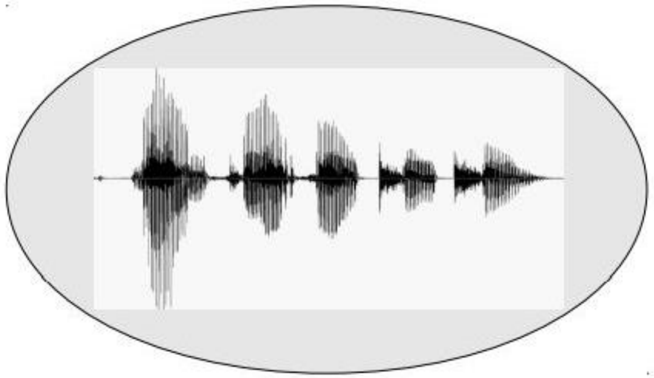
Applications

- Speech Synthesis
- Parsing
- Machine Translation
- Information Extraction

Applications

- Speech Synthesis

How to produce 'lead'?

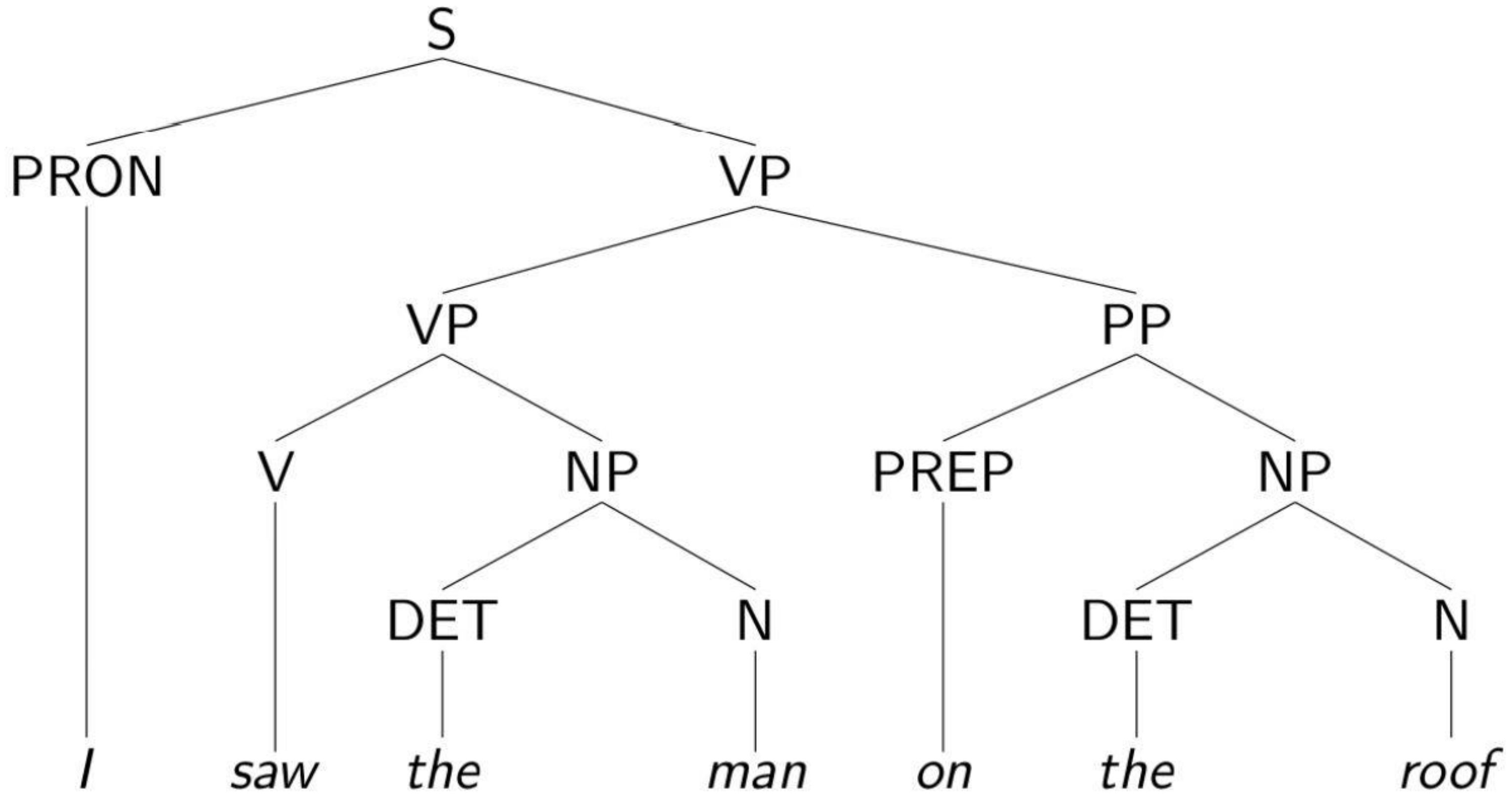


Applications

- Machine Translation

Applications

- Parsing



POS Tagset

- There are so many parts of speech tagsets we can draw
- Choosing a standard tagset is essential
- Tag types
  - Coarse-grained
    - noun
    - verb
    - adjective
    - ...
- Fine-grained
  - noun-proper-singular, noun-proper-plural, noun-common-mass, ..
  - verb-past, verb-present-3rd, verb-base, ...
  - adjective-simple, adjective-comparative, ...
  - ...

Penn TreeBank 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*			

"Universal Dependencies" Tagset

	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*, *yesterday*
	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..	*Regina*, *IBM*, *Colorado*
	INTJ	Interjection: exclamation, greeting, yes/no response, etc.	*oh*, *um*, *yes*, *hello*
Closed Class Words	**ADP**	Adposition (Preposition/Postposition): marks a noun's spacial, temporal, or other relation	*in, on, by under*
	AUX	Auxiliary: helping verb marking tense, aspect, mood, etc.,	*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*
	PART	Particle: a preposition-like form used together with a verb	*up, down, on, off, in, out, 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

Part-of-Speech Tagging

- Definition
  - The process of assigning a part of speech to each word in a text

- Challenge
  - Words often have more than one POS

On my back
The back door
Pay the money back
Promised to back the bill

Part-of-Speech Tagging

- Definition
  - The process of assigning a part of speech to each word in a text

- Challenge
  - Words often have more than one POS

On my back[NN]
The back[JJ] door
Pay the money back[RB]
Promised to back[VB] the bill

Distribution of Ambiguities

		45-tag Treebank Brown	
Unambiguous (1 tag)		**38,857**	
Ambiguous (2–7 tags)		**8844**	
Details:	2 tags	6,731	
	3 tags	1621	
	4 tags	357	
	5 tags	90	
	6 tags	32	
	7 tags	6	(*well*, *set*, *round*, *open*, *fit*, *down*)
	8 tags	4	(*'s*, *half*, *back*, *a*)
	9 tags	3	(*that*, *more*, *in*)

Distribution of Ambiguities

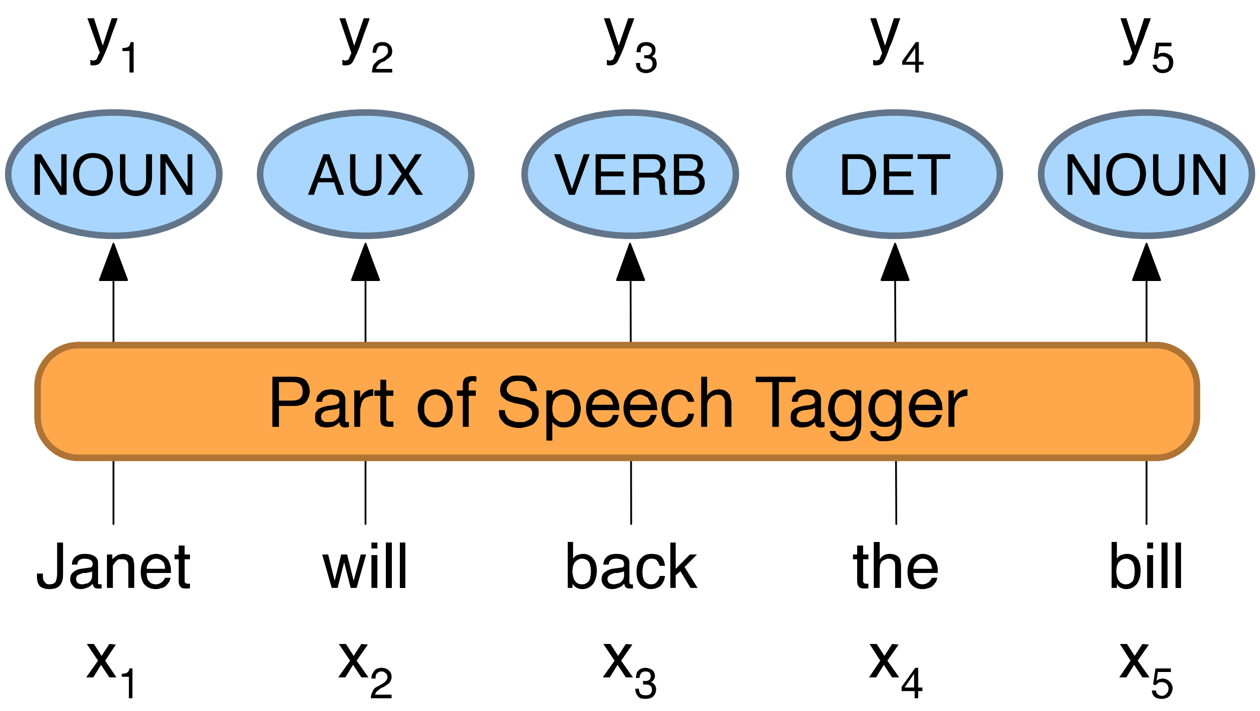
- The frequency of ambiguous words are relatively high
  - 11.5% of word types
  - 40% of word tokens

Goal

- Using a set of labeled data to train a model
- Using the trained model to predict the POS tag of the unseen words

Part-of-Speech Tagging

- Map from sequence $x_1,\ldots,x_n$ of words to $y_1,\ldots,y_n$ of POS tags



POS Tagging

Plays well with others

Plays	NNS/VBZ
well	UH/JJ/NN/RB
with	IN
others	NNS

Plays[VBZ] well[RB] with[IN] others[NNS]

Basic Models

- Baseline model
  - Tagging unambiguous words with the correct label
  - Tagging ambiguous words with their most frequent label
  - Tagging unknown words as a noun

Already performs around 90%

- Basic classification model
  - Classifying each word to the pre-define list of POS tags

Outline

- Parts of Speech Tagging
- **Named Entity Recognition**
- Sequence Modeling
- PGM-based Model for Sequence Labelling
- Evaluation

Named Entities

- **Named entity**, in its core usage, means anything that can be referred to with a proper name. Most common 4 tags:
  - PER (Person): “Marie Curie”
  - LOC (Location): “New York City”
  - ORG (Organization): “Stanford University”
  - GPE (Geo-Political Entity): "Boulder, Colorado"

- Often multi-word phrases
  - But the term is also extended to things that aren't entities

Named Entities

- Dates and times
- Prices
- Measure (Percent, Money, Weight, ...)
- Religious
- Book title
- Movie title
- Drug name

Named Entity Recognition

- Also known as named entity tagging

- The task of named entity recognition (NER):
  - find spans of text that constitute proper names
  - tag the type of the entity.

NER output

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**].

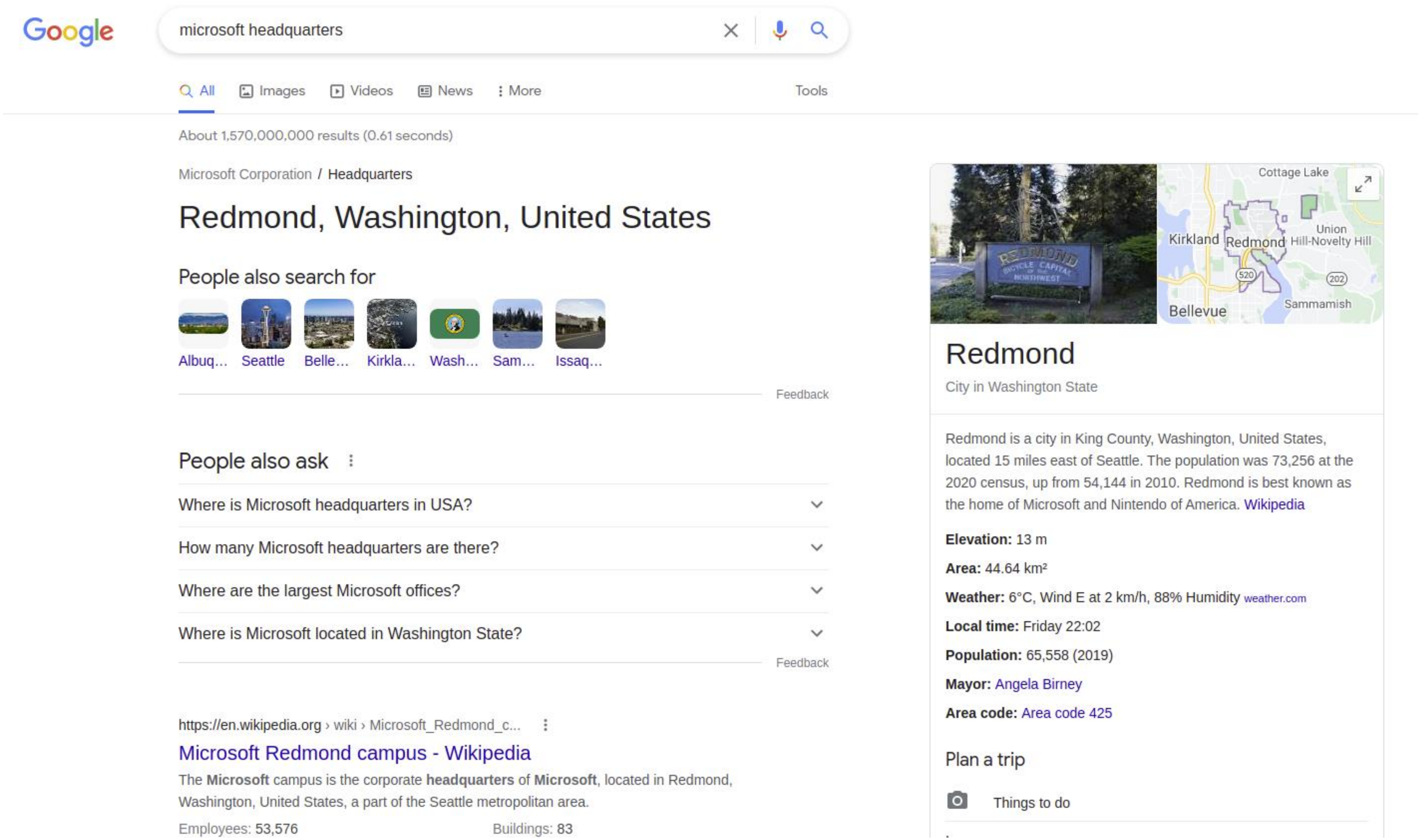
Motivation

- Sentiment analysis: consumer's sentiment toward a particular company or person?
- Factual information and knowledge are normally expressed by named entities
  - Who, Whom, Where, When, ...
- Question answering systems are looking for named entities to answer users' questions
- Named entity recognition is the core of the information extraction systems

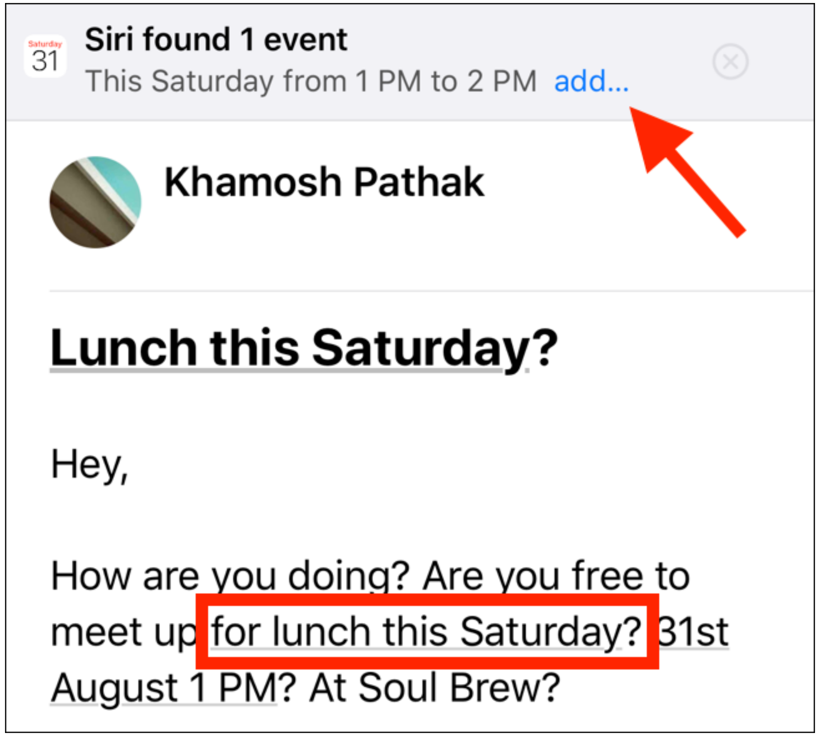
Applications

- Finding the important information of an event from an invitation
  - Date, Time, Location, Host, Contact person
- Finding the main information of a company from its reports
  - Founder, Board members, Headquarters, Profits
- Finding medical information from medical literature
  - Drugs, Genes, Interaction products
- Finding the target of sentiments
  - Products, Celebrities

Applications



Applications



Why NER is hard

- Segmentation
  - In POS tagging, no segmentation problem since each word gets one tag.
  - In NER we have to find and segment the entities!

- Type ambiguity

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

BIO Tagging

- How can we turn this structured problem into a sequence problem like POS tagging, with one label per word?

- [PER Jane Villanueva] of [ORG United] , a unit of [ORG United Airlines Holding] , said the fare applies to the [LOC Chicago] route.

BIO Tagging

- [PER Jane Villanueva] of [ORG United] , a unit of [ORG United Airlines Holding] , said the fare applies to the [LOC Chicago] route.

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

- Now we have one tag per token!!!

BIO Tagging

- B: token that *begins* a span
- I: tokens *inside* a span
- O: tokens outside of any span

- # of tags (where n is #entity types):
- 1 O tag,
- *n* B tags,
- *n* I tags
- Total: *2n+1*

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

BIO Tagging variants: IO and BIOES

- [PER Jane Villanueva] of [ORG United] , a unit of [ORG United Airlines Holding] , said the fare applies to the [LOC Chicago] route.

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

IO vs. IOB Encoding

John	PER
Shows	O
Mary	PER
Hermann	PER
Hesse	PER
’s	O
book	O
.	O

John	B-PER
Shows	O
Mary	B-PER
Hermann	B-PER
Hesse	I-PER
’s	O
book	O
.	O

- Although IOB is more accurate, some systems prefer IO for the following reasons
  - IO is much faster than IOB
  - The above case happens very rarely. Even in such cases achieving correct results with IOB is difficult and unlikely

Basic Models

- List lookup
  - Using dictionary of named entities

- Learning models
  - Similar to POS tagging

List lookup

- Extensive list of names are available via various resources
- The name lists include lists of
  - Entities
    - Organization, government, airline, educational, ..
    - Location, continent, country, state, city, ...
    - Person first name, last name, ...
  - Entity cues
    - Typical words in organization; e.g., “Limited" or “Incorporated"
    - Person title; e.g., “Mister", “Lord"
- The terms “gazetteer", “lexicon" and “dictionary" are often used interchangeably with the term “list"
  - Gazetteer originally referred to a large list of place names but it became a more general terminology in the NER task

Outline

- Parts of Speech Tagging
- Named Entity Recognition
- **Sequence Modeling**
- PGM-based Model for Sequence Labelling
- Evaluation

Sequence Labelling

- Similar to a normal classification task
  - Feature Selection
  - Algorithm

POS Tagging with Word Features

- Features

Word	the: the → DT
Prefixes	unbelievable: un- → JJ
Suffixes	slowly: -ly → RB
Lowercased word	Importantly: importantly → RB
Capitalization	Stefan: [CAP] → NNP
Word shapes	35-year: d-x → JJ

- Model
  - Maximum Entropy P(t|w)

Data	Performance
Overall	93.7
Unknown	82.6

NER with Word Features

- Features

Word	Germany: Germany
POS tag	Washington: NNP
Capitalization	Stefan: [CAP]
Punctuation	St.: [PUNC]
Lowercased word	Book: book
Suffixes	Spanish: -ish
Word shapes	1920-2008: dddd-dddd

- List lookup

POS Tagging

- More Features?

They$_{[PRP]}$ left$_{[VBD]}$ as$_{[IN]}$ soon$_{[RB]}$ as$_{[IN]}$ he$_{[PRP]}$ arrived$_{[NBD]}$.

- Better Algorithm
  - Using Sequence Modeling

Sequence Modeling

- Many of the NLP techniques should deal with data represented as sequence of items
  - Characters, Words, Phrases, Lines, ...

警察枪杀了那个逃

B I BI B B B B I

Sequence Modeling

- Many of the NLP techniques should deal with data represented as sequence of items
  - Characters, Words, Phrases, Lines, ...

警察枪杀了那个逃

B I BI B B B B I

$I_{[PRP]}$ $saw_{[VBP]}$ $the_{[DT]}$ $man_{[NN]}$ $on_{[IN]}$ $the_{[DT]}$ $roof_{[NN]}$.

Sequence Modeling

- Many of the NLP techniques should deal with data represented as sequence of items
  - Characters, Words, Phrases, Lines, ...

警察枪杀了那个逃

B I BI B B B B I

$I_{[PRP]}$ $saw_{[VBP]}$ $the_{[DT]}$ $man_{[NN]}$ $on_{[IN]}$ $the_{[DT]}$ $roof_{[NN]}$.

Steven	Paul	Jobs,	co-founder	of	Apple	Inc,	was	born	in	California.
PER	PER	PER	O	O	ORG	ORG	O	O	O	LOC

Sequence Modeling

- Two types of information
  - Local
  - Contextual

Sequence Modeling

- Making a decision based on the
  - Current Observation
    - Word (W0)
    - Prefix
    - Suffix
    - Lowercased word
    - Capitalization
    - Word shape
  - Surrounding observations
    - W_{+1}
    - W_{-1}
  - Previous decisions
    - T_{-1}
    - T_{-2}

Methods for Sequence Labeling

- PGM-based Methods
  - Hidden Markov Model (HMM)
  - Maximum Entropy Markov Model (MEMM)
  - Conditional Random Fields (CRF)

** These are all classifiers (i.e., supervised learning) which model sequences (rather than individual random variables)*

- Neural Methods
  - Recurrent Neural Networks (RNN)
  - Transformers

Outline

- Parts of Speech Tagging
- Named Entity Recognition
- Sequence Modeling
- **PGM-based Model for Sequence Labelling**
- Evaluation

Hidden Markov Model (HMM)

- Finding the best sequence of tags ($t_1 \ldots t_n$) that corresponds to the sequence of observations ($w_1 \ldots w_n$)
- Probabilistic View
  - Considering all possible sequences of tags
  - Choosing the tag sequence from this universe of sequences, which is most probable given the observation sequence

$$\hat{t}_1^n = \operatorname{argmax}_{t_1^n} P(t_1^n \mid w_1^n)$$

Using Bayes Rule

$$\hat{t}_1^n = \operatorname{argmax}_{t_1^n} P(t_1^n \mid w_1^n)$$

$$P(A \mid B) = \frac{P(B \mid A).P(A)}{P(B)}$$

$$P(t_1^n \mid w_1^n) = \frac{P(w_1^n \mid t_1^n).P(t_1^n)}{P(w_1^n)}$$

$$\hat{t}_1^n = \operatorname{argmax}_{t_1^n} P(w_1^n \mid t_1^n).P(t_1^n)$$

Using Markov Assumption

$$\hat{t}_1^n = \text{argmax}_{t_1^n} P(w_1^n \mid t_1^n).P(t_1^n)$$

$$P(w_1^n \mid t_1^n) \approx \prod_{i=1}^{n} P(w_i \mid t_i)$$

$$P(t_1^n) \approx \prod_{i=1}^{n} P(t_i \mid t_{i-1})$$

$$\hat{t}_1^n = \text{argmax}_{t_1^n} \prod_{i=1}^{n} P(w_i \mid t_i).P(t_i \mid t_{i-1})$$

Two Probabilities

- The tag transition probabilities: $P(t_i \mid t_{i-1})$
  - Finding the likelihood of a tag to proceed by another tag
  - Similar to the normal bigram model

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

Two Probabilities

- The word likelihood probabilities: $P(w_i \mid t_i)$
  - Finding the likelihood of a word to appear given a tag

$$P(w_i|t_i) = \frac{\#(t_i, w_i)}{\#(t_i)}$$

Two Probabilities

- Zero probability problem Solution:
  - similar to language modelling, use the smoothing method for both probabilities

Two Probabilities

I$_{[PRP]}$ saw$_{[VBP]}$ the$_{[DT]}$ man$_{[NN?]}$ on$_{[]}$ the$_{[]}$ roof$_{[]}$.

$$P([NN] \mid [DT]) = \frac{C([DT],[NN])}{C([DT])}$$

$$P(man \mid [NN]) = \frac{C([NN], man)}{C([NN])}$$

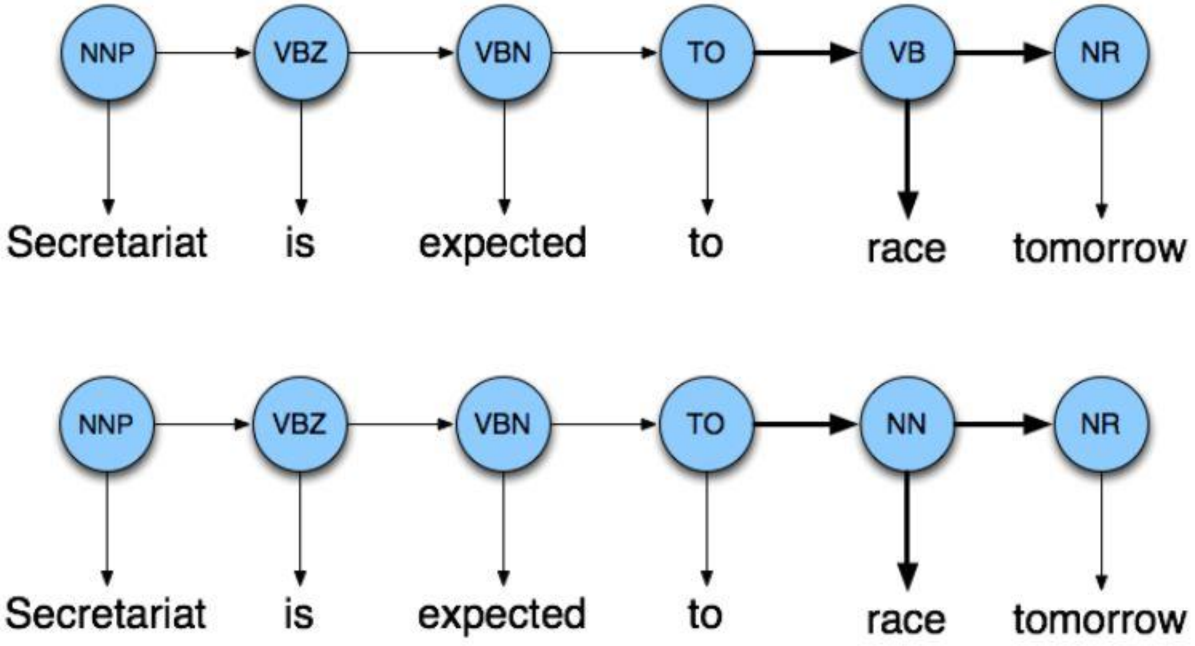
Ambiguity

Secretariat$_{[NNP]}$ is$_{[VBZ]}$ expected$_{[VBN]}$ to$_{[TO]}$ race$_{[VB]}$ tomorrow$_{[NR]}$.

People$_{[NNS]}$ inquire$_{[VB]}$ the$_{[DT]}$ reason$_{[NN]}$ for$_{[IN]}$ the$_{[DT]}$ race$_{[NN]}$.

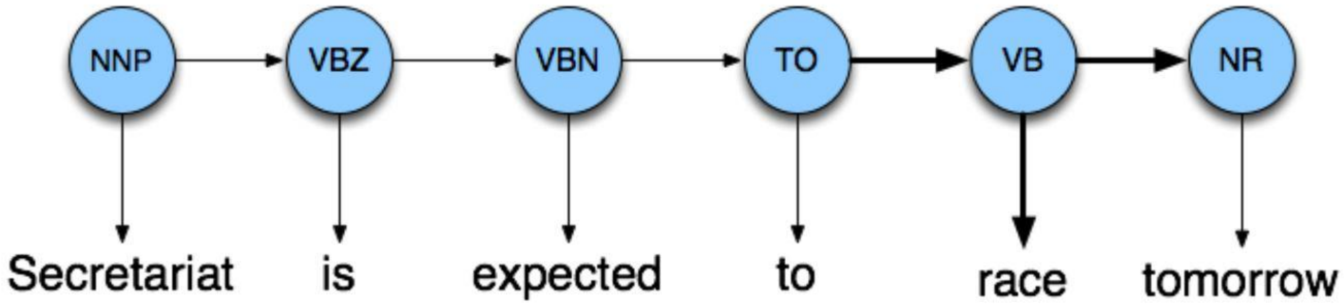
Ambiguity

Secretariat$_{[NNP]}$ is$_{[VBZ]}$ expected$_{[VBN]}$ to$_{[TO]}$ race$_{[VB]}$ tomorrow$_{[NR]}$.



Ambiguity

Secretariat[NNP] is[VBZ] expected[VBN] to[TO] **race**[VB] tomorrow[NR] .



$$P(VB \mid TO) = 0.83$$

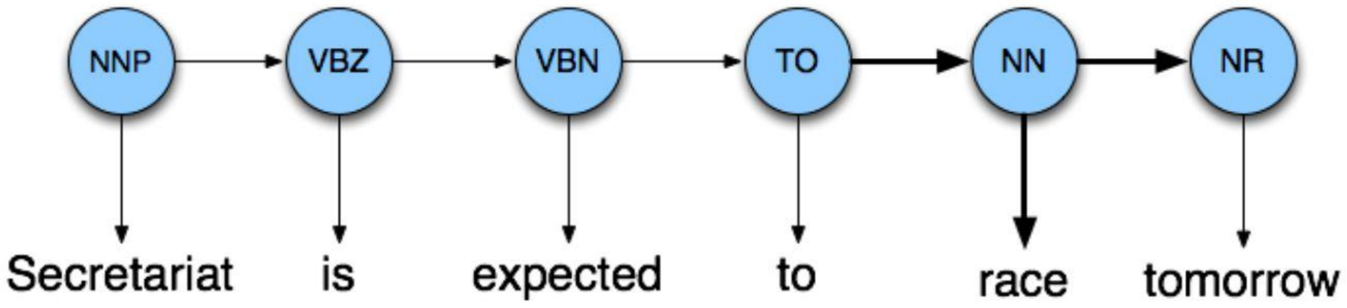
$$P(race \mid VB) = 0.00012$$

$$P(NR \mid VB) = 0.0027$$

$$P(VB \mid TO)P(NR \mid VB)P(race \mid VB) = 0.00000027$$

Ambiguity

Secretariat$_{[NNP]}$ is$_{[VBZ]}$ expected$_{[VBN]}$ to$_{[TO]}$ **race**$_{[VB]}$ tomorrow$_{[NR]}$.



$$P(NN \mid TO) = 0.00047$$

$$P(race \mid NN) = 0.00057$$

$$P(NR \mid NN) = 0.0012$$

$$P(NN \mid TO)\ P(NR \mid NN)\ P(race \mid NN) = 0.00000000032$$

Performance

- Model
  - Maximum Entropy P(t|w)

Data	Performance
Overall	93.7
Unknown	82.6

  - HMM

Data	Performance
Overall	96.2
Unknown	86.0

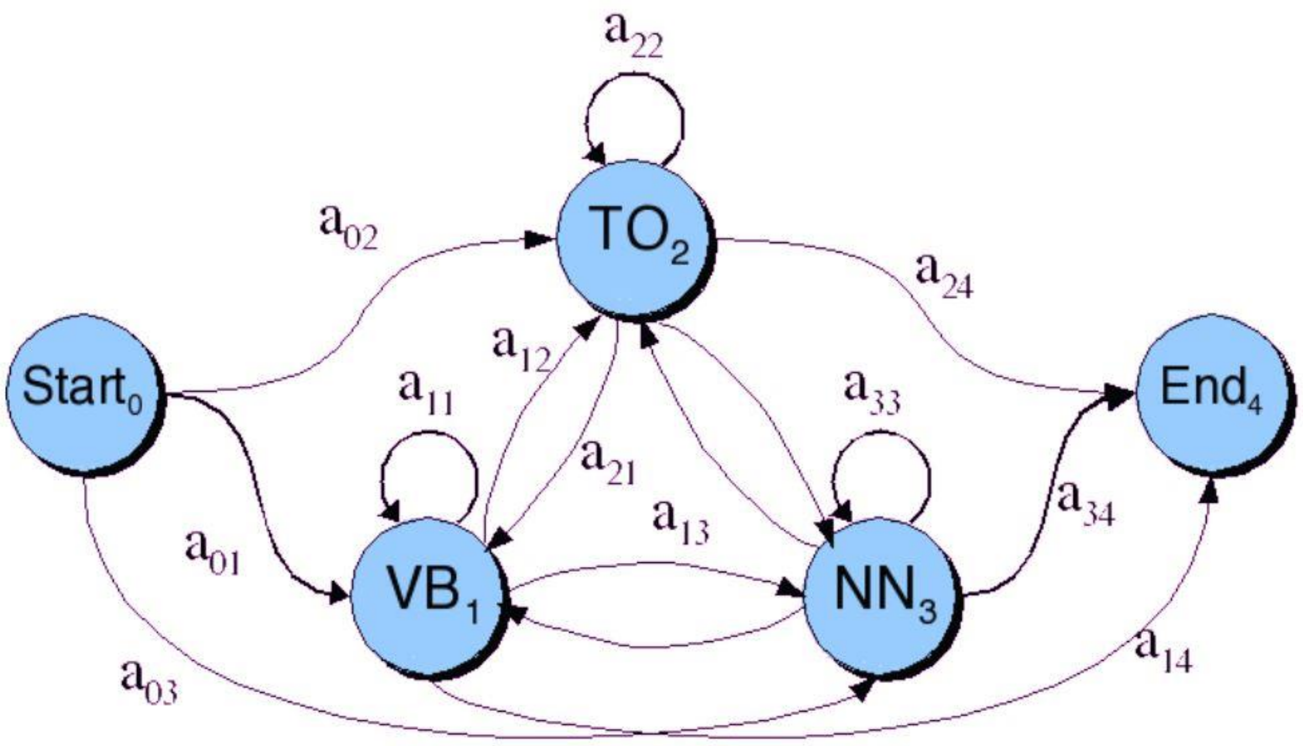
- Upper bound (human agreement): ~98%

Hidden Markov Model (HMM)

- A weighted finite-state automaton adds probabilities to the arcs
  - The probabilities leaving any arc must sum to one
- An HMM is an extension of a Markov chain in which the input symbols are not the same as the states
- We do not know which state we are in
  - The output symbols are words
  - The hidden states are POS tags

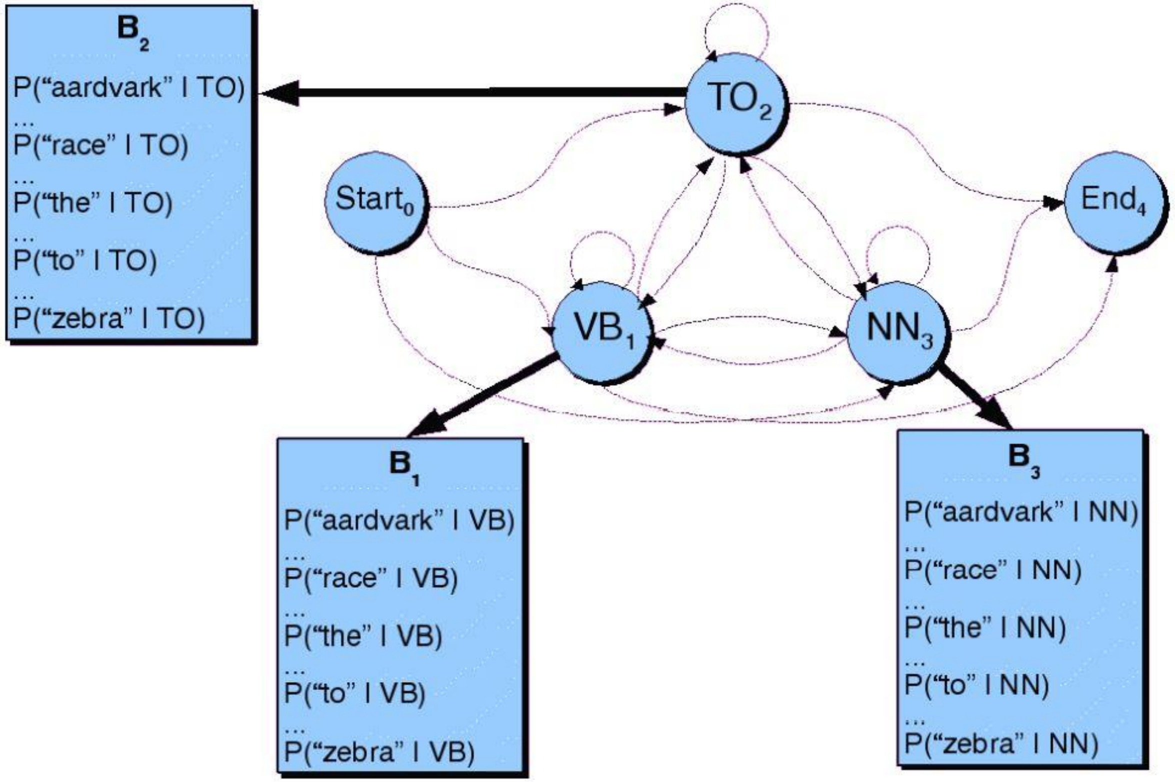
Hidden Markov Model (HMM)

- Transition probabilities



Hidden Markov Model (HMM)

- Word likelihood probabilities



The Viterbi Algorithm

- Viterbi inference
  - Memorizing the model using dynamic programming
  - Considering the small window of previous decisions

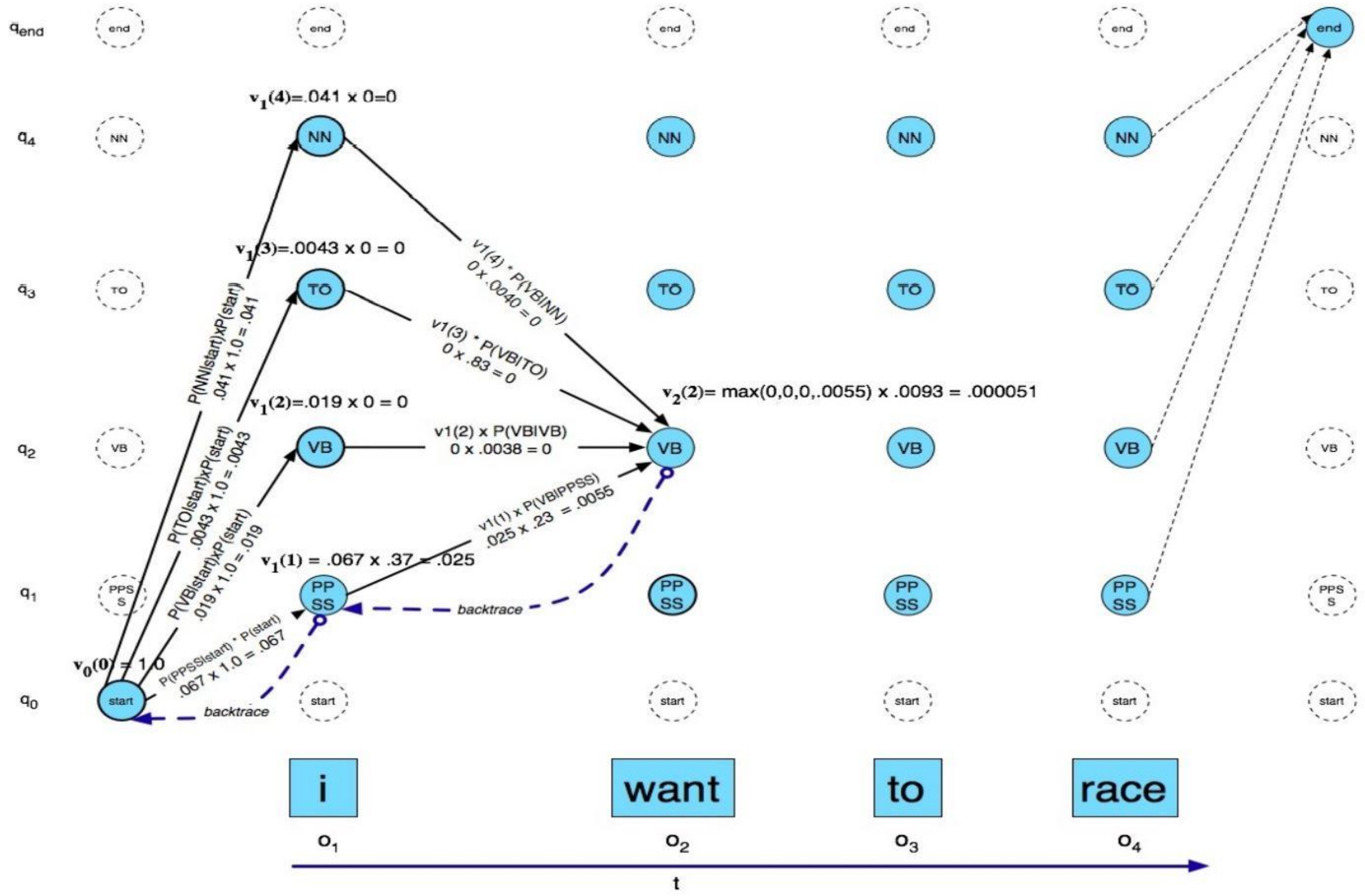
The Viterbi Algorithm

- Creating an array
  - Columns corresponding to inputs
  - Rows corresponding to possible states
- Sweeping through the array in one pass filling the columns left to right using the transition probabilities and observation probabilities
- Storing the max probability path to each cell (not all paths) using dynamic programming

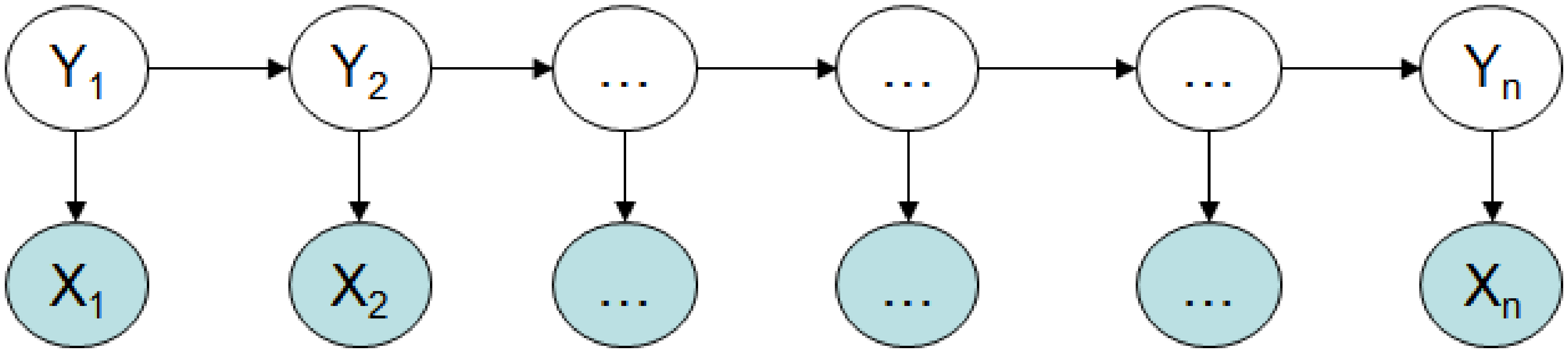
The Viterbi Algorithm

- Basic idea behind the algorithm: the recursive definition for finding the maximum probability

The Viterbi Algorithm



Hidden Markov Model (HMM)



Hidden Markov Model (HMM)

- Advantages
  - A strong statistical foundation with efficient learning algorithms where learning can take place directly from raw sequence data.
  - can perform a wide variety of operations and algorithms

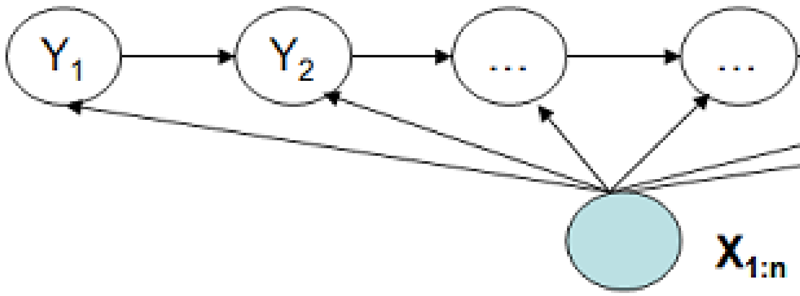
- Disadvantages
  - HMM is only dependent on every state and its corresponding observed object

  (The sequence labeling, in addition to having a relationship with individual words, also relates to such aspects as the observed sequence length, word context and others.)
  - The target function and the predicted target function do not match

  (HMM acquires the joint distribution $P(Y, X)$ of the state and the observed sequence, while in the estimation issue, we need a conditional probability $P(Y|X)$)

Maximum Entropy Markov Model (MEMM)

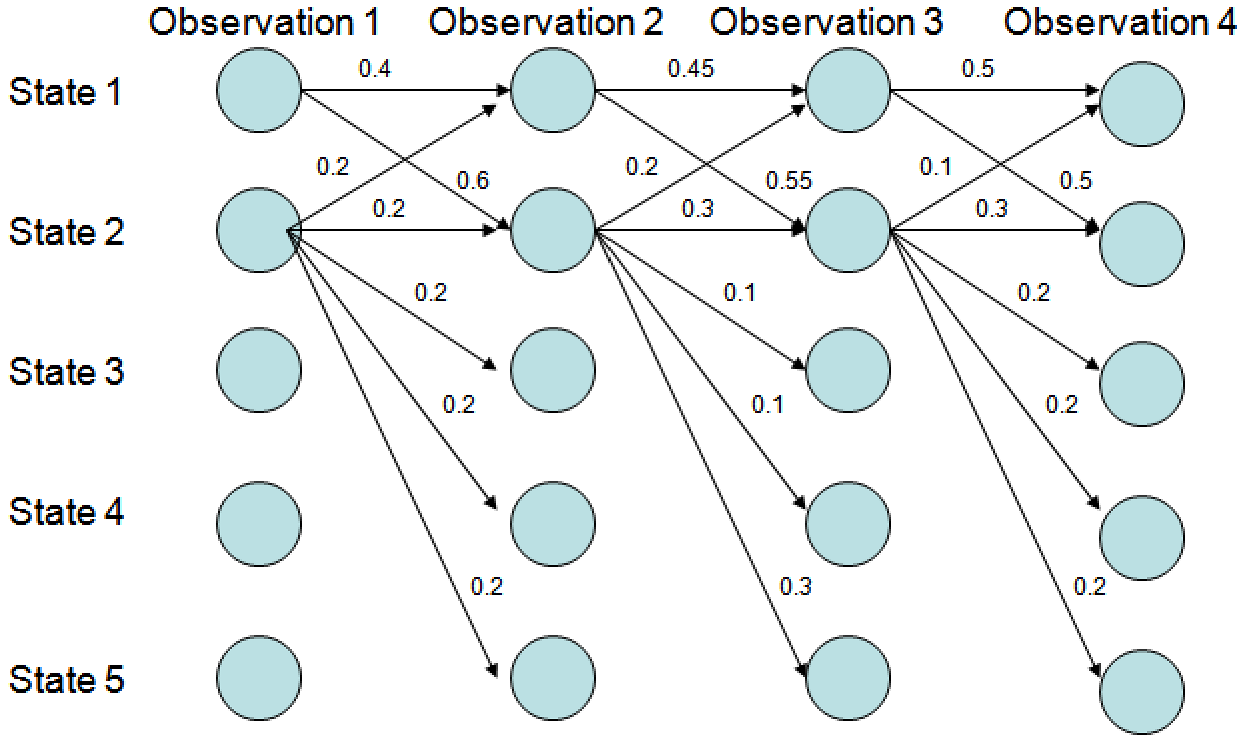
- Also known as Conditional Markov Model (CMM)
- The classifier decision is conditioned on the evidence from observations and previous decisions
- MEMM takes into account the dependencies between neighboring states and the entire observed sequence, hence a better expression ability. MEMM does not consider P(X), which reduces the modeling workload and learns the consistency between the target function and the estimated function.



$$P(\mathbf{y}_{1:n}|\mathbf{x}_{1:n}) = \prod_{i=1}^{n} P(y_i|y_{i-1},\mathbf{x}_{1:n}) = \prod_{i=1}^{n} \frac{\exp(\mathbf{w}^T\mathbf{f}(y_i,y_{i-1},\mathbf{x}_{1:n}))}{Z(y_{i-1},\mathbf{x}_{1:n})}$$

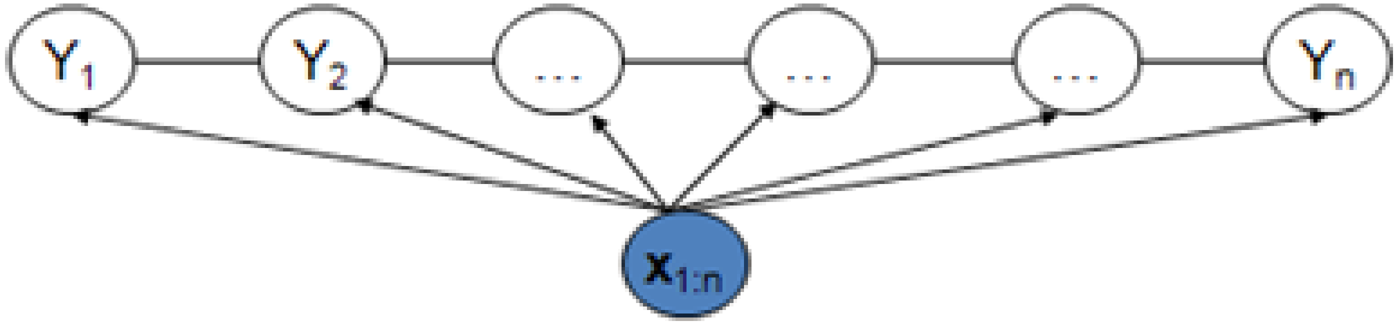
Maximum Entropy Markov Model (MEMM)

- Label Bias Problem in MEMM
  - Most Likely Path: 1 → 1 → 1 → 1
  - Although locally it seems state 1 wants to go to state 2 and state 2 wants to remain in state 2 or 5.
  - Preference of states with lower number of transitions over others



Conditional Random Field (CRF)

- Another alternative for sequence modeling
- A whole-sequence of labels (classes) is conditioned to the whole-sequence of data items rather than a chaining of local models
  - The space of c' is now the space of sequences



Conditional Random Field (CRF)

- Advantages
  - CRF addresses the labeling bias problem of MEMM
    - MEMM adopts local variance normalization while CRF adopts global variance normalization
  - CRF does not have as strict independence assumptions as HMM does, it can accommodate any context information

- Disadvantages
  - CRF is computationally complex at the training stage of the algorithm

Outline

- Parts of Speech Tagging
- Named Entity Recognition
- Sequence Modeling
- PGM-based Model for Sequence Labelling
- **Evaluation**

Evaluating POS Tagging

- Comparing the output of a tagger with a human-labelled gold standard
- Accuracy:

$$\text{Accuracy} = \frac{\#\text{currectly taggd words}}{\#\text{total word token}}$$

Evaluating POS Tagging

- Accuracy:

$$\text{Accuracy} = \frac{\text{tp}}{\text{N}}$$

$$\text{Accuracy} = \frac{\sum_{c}^{C} tp_c}{\text{N}}$$

Evaluating POS Tagging

- The accuracy score doesn’t show everything
- It is useful to know what is misclassified as what
- Solution: providing a confusion matrix
  - A matrix (# tags x #tags): the rows correspond to the correct tags and the columns correspond to the tagger output
  - Cell(i, j) gives the count of the number of times tag i was classified as tag j
  - The leading diagonal elements correspond to correct classifications
  - Off diagonal elements correspond to misclassifications
- A good approach for error analysis

Evaluating NER

I. Surface string and entity type match

Golden Standard		System Prediction	
Surface String	Entity Type	Surface String	Entity Type
in	O	in	O
New	B-LOC	New	B-LOC
York	I-LOC	York	I-LOC
.	O	.	O

Evaluating NER

II. System hypothesized an entity

Golden Standard		System Prediction	
Surface String	**Entity Type**	**Surface String**	**Entity Type**
an	O	an	O
Awful	O	Awful	B-ORG
Headache	O	Headache	I-ORG
in	O	in	O

III. System misses an entity

Golden Standard		System Prediction	
Surface String	**Entity Type**	**Surface String**	**Entity Type**
in	O	in	O
Palo	B-LOC	Palo	O
Alto	I-LOC	Alto	O
,	O	,	O

Evaluating NER

IV. System assigns the wrong entity type

Golden Standard		System Prediction	
Surface String	**Entity Type**	**Surface String**	**Entity Type**
I	O	I	O
live	O	live	O
in	O	in	O
Palo	B-LOC	Palo	B-ORG
Alto	I-LOC	Alto	I-ORG
,	O	,	O

V. System gets the boundaries of the surface string wrong

Golden Standard		System Prediction	
Surface String	**Entity Type**	**Surface String**	**Entity Type**
Unless	O	Unless	B-PER
Karl	B-PER	Karl	I-PER
Smith	I-PER	Smith	I-PER
resigns	O	resigns	O

VI. System gets the boundaries and entity type wrong

Golden Standard		System Prediction	
Surface String	**Entity Type**	**Surface String**	**Entity Type**
Unless	O	Unless	B-ORG
Karl	B-PER	Karl	I-ORG
Smith	I-PER	Smith	I-ORG
resigns	O	resigns	O

Evaluating NER

- F1-Score at token level (word level)
- F1-Score at entity level (phrase/segment level)

Evaluating NER

- Example of entity level F1-score

TRUTH:

Michael Kearns and Sebastian Seung will start Monday's tutorial, followed by Richard M. Karpe and Martin Cooke.

PRED:

Michael Kearns and Sebastian Seung will start Monday's tutorial, followed by Richard M. Karpe and Martin Cooke.

$$\text{Precision} = \frac{\text{\# correctly predicted segments}}{\text{\# predicted segments}} = \frac{2}{6}$$

$$\text{Recall} = \frac{\text{\# correctly predicted segments}}{\text{\# true segments}} = \frac{2}{4}$$

$$\text{F1} = \text{Harmonic mean of Precision \& Recall} = \frac{1}{((1/P) + (1/R)) / 2}$$

Further Reading

- Speech and Language Processing (3rd ed. draft)
  - Chapter 8