Natural Language Processing
SoSe 2014

Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Part-of-speech tagging and named-entity recognition

Dr. Mariana Neves

May 7th, 2014

(based on the slides of Dr. Saeedeh Momtazi)

Outline

- Part of Speech Tagging
- Named Entity Recognition
- Sequential Modeling

Outline

- Part of Speech Tagging
- Named Entity Recognition
- Sequential Modeling

Parts of speech (POS)

- 8 Parts of speech are traditionally used to summarize 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, nature, 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 (limited number of words, do not grow usually)
  - Determiners: the, some, a, an, ...
  - Pronouns: she, he, I, ...
  - Prepositions: to, in, on, under, over, by, ...
  - Auxiliaries: can, should, have, had, ...
  - Conjunctions: and, or
  - Particles: off, up
  - Interjections: Ow, Eh
- Open (unlimited number of words)
  - Nouns
  - Verbs
  - Adjectives
  - Adverbs

Applications

- Speech Synthesis
- Parsing
- Machine Translation
- Information Extraction

Applications

- Speech Synthesis
  - „content“
    - „Eggs have a high protein content.“
    - „She was content to step down after four years as chief executive.“

http://www.thefreedictionary.com/content

Applications

- Machine Translation
  - „I like ...“
    - „Ich mag ....“
    - „Ich wie ...“

Applications

- Parsing

Your query

I saw the man on the roof

Tagging

```
I/PRP  saw/VBD  the/DT  man/NN  on/IN  the/DT  roof/NN
```

Parse

```
(ROOT
  (S
    (NP (PRP I))
    (VP (VBD saw)
      (NP (DT the) (NN man))
      (PP (IN on)
        (NP (DT the) (NN roof))))))
```

http://nlp.stanford.edu:8080/parser/index.jsp

Applications

- Information Extraction

```
> echo "Inhibition of NF-kappaB activation reversed the anti-apoptotic effect of isochamaejasmin." | ./geniatagger

Inhibition      Inhibition      NN      B-NP    O
of              of              IN      B-PP    O
NF-kappaB       NF-kappaB       NN      B-NP    B-protein
activation      activation      NN      I-NP    O
reversed        reverse         VBD     B-VP    O
the             the             DT      B-NP    O
anti-apoptotic  anti-apoptotic  JJ      I-NP    O
effect          effect          NN      I-NP    O
of              of              IN      B-PP    O
isochamaejasmin isochamaejasmin NN      B-NP    O
.               .               .       O       O
```

http://www.nactem.ac.uk/tsujii/GENIA/tagger/

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, ...

POS Tagset

- Brown tagset
  - Brown corpus
  - 87 tags
- C5 tagset
  - 61 tags
- C7 tagset
  - 146 tags!!
- Penn TreeBank
  - A large annotated corpus of English tagset: 45 tags

Penn TreeBank Tagset

POS Tag	Description	Example
CC	coordinating conjunction	and
CD	cardinal number	1, third
DT	determiner	the
EX	existential there	*there* is
FW	foreign word	d'hoevre
IN	preposition/subordinating conjunction	in, of, like
JJ	adjective	green
JJR	adjective, comparative	greener
JJS	adjective, superlative	greenest
LS	list marker	1)
MD	modal	could, will
NN	noun, singular or mass	table
NNS	noun plural	tables
NNP	proper noun, singular	John
NNPS	proper noun, plural	Vikings
PDT	predeterminer	*both* the boys
POS	possessive ending	friend's
PRP	personal pronoun	I, he, it
PRP$	possessive pronoun	my, his
RB	adverb	however, usually, naturally, here, good
RBR	adverb, comparative	better
RBS	adverb, superlative	best
RP	particle	give *up*
TO	to	*to* go, *to* him
UH	interjection	uhhuhhuhh
VB	verb, base form	take
VBD	verb, past tense	took
VBG	verb, gerund/present participle	taking
VBN	verb, past participle	taken
VBP	verb, sing. present, non-3d	take
VBZ	verb, 3rd person sing. present	takes
WDT	wh-determiner	which
WP	wh-pronoun	who, what
WP$	possessive wh-pronoun	whose
WRB	wh-abverb	where, when

http://www.americannationalcorpus.org/OANC/penn.html

POS 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
  - Win the voters back$_{[RB]}$
  - Promised to back$_{[VB]}$ the bill

Distribution of Ambiguities

- 45-tags Brown corpus (word types)
  - Unambiguous (1 tag): 38,857
  - Ambiguous: 8,844
    - 2 tags: 6,731
    - 3 tags: 1,621
    - 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)

POS Tagging

- Plays well with others

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

- $\text{Plays}_{\text{[VBZ]}}$ $\text{well}_{\text{[RB]}}$ $\text{with}_{\text{[IN]}}$ $\text{others}_{\text{[NNS]}}$

Performance

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

- Performs around 90%

Outline

- Part of Speech Tagging
- Named Entity Recognition
- Sequential Modeling

Motivation

- Factual information and knowledge are normally expressed by named entities
  - Who, Whom, Where, When, Which, ...

- 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 information from biomedical literature
  - Drugs, Genes, Interaction products
- Finding the target of sentiments
  - Products, Celebrities

Applications

TREAT DIS
28 OBJECTIVE : To study the safety and efficacy of methylphenidate in children with the dual diagnosis of epilepsy and

DIS
attention deficit hyperactivity disorder (ADHD) .

DISONLY DISONLY
29 STUDY DESIGN : Thirty children , aged 6.4 to 16.4 years , with epilepsy and ADHD were studied during a 4-month period .

TREATONLY
30 During the initial 2 months of the study , the children were treated with antiepileptic drugs (AEDs) only , and for the remaining 2

TREATONLY
months , methylphenidate was added at a morning dose of 0.3 mg/kg .

31 They underwent neurologic assessment , brain computed tomography , IQ testing , and assessment with the Childhood Behavior

TREATONLY
Checklist at baseline before methylphenidate therapy .

32 Electroencephalography , AED determinations , and the continuous-performance task (CPT) test were done at baseline and after 2 months of methylphenidate therapy .

TREATONLY
33 A double-blind , crossover design was used to compare the effects of methylphenidate versus placebo on an electroencephalogram , AED levels , and the CPT .

TREATONLY TREATONLY
34 On the 2 days of testing , the child received AEDs and a capsule containing either placebo or methylphenidate .

35 RESULTS : None of the 25 children of this sample who were seizure free had attacks while taking methylphenidate .

DISONLY
36 Of the 5 children with seizures , 3 had an increase in attacks , whereas the other 2 showed no change or a reduction .

37 There were no significant changes in AED levels or electroencephalographic findings .

Named Entity Recognition (NER)

- Finding named entities in a text
- Classifying them to the corresponding classes

- „Steven Paul Jobs, co-founder of Apple Inc, was born in California."
- „Steven Paul Jobs, co-founder of Apple Inc, was born in California."
- „Steven Paul Jobs [PER], co-founder of Apple Inc [ORG], was born in California [LOC]."

Named Entity Classes

- Person
  - Person names
- Organization
  - Companies, Government, Organizations, Committees, ..
- Location
  - Cities, Countries, Rivers, ..
- Date and time expression
- Measure
  - Percent, Money, Weight, ...
- Book, journal title
- Movie title
- Gene, disease, drug name

Named Entity Classes (IO)

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

Named Entity Classes (BIO/IOB)

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

Named Entity Classes (BIEWO)

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

NER Ambiguity (IO vs. IOB encoding)

John	PER	John	B-PER
shows	O	shows	O
Mary	PER	Mary	B-PER
Hermann	PER	Hermann	B-PER
Hesse	PER	Hesse	I-PER
's	O	's	O
book	O	book	O
.	O	.	O

NER Ambiguity

- Ambiguity between named entities and common words
  - May: month, verb, surname
  - Genes: VIP, hedgehog, deafness, wasp, was, if

- Ambiguity between named entity types
  - Washington (Location or Person)

Outline

- Part of Speech Tagging
- Named Entity Recognition
- Sequential Modeling

Task

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

POS Tagging

- 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

POS Tagging

- Model
  - Maximum Entropy: P(t|w)
    - Overall words: 93.7%
    - Unknown words: 82.6%

NER

- 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

NER

- List lookup
  - Extensive list of names are available via various resources
  - Gazetteer: a large list of place names
  - Biomedical: database of genes, proteins, drugs names
  - Usually good precision, but low recall (variations)

POS Tagging

- More Features?

  They[PRP] left[VBD] as[IN] soon[RB] as[IN] he[PRP] arrivied[VBD]

- 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, ...

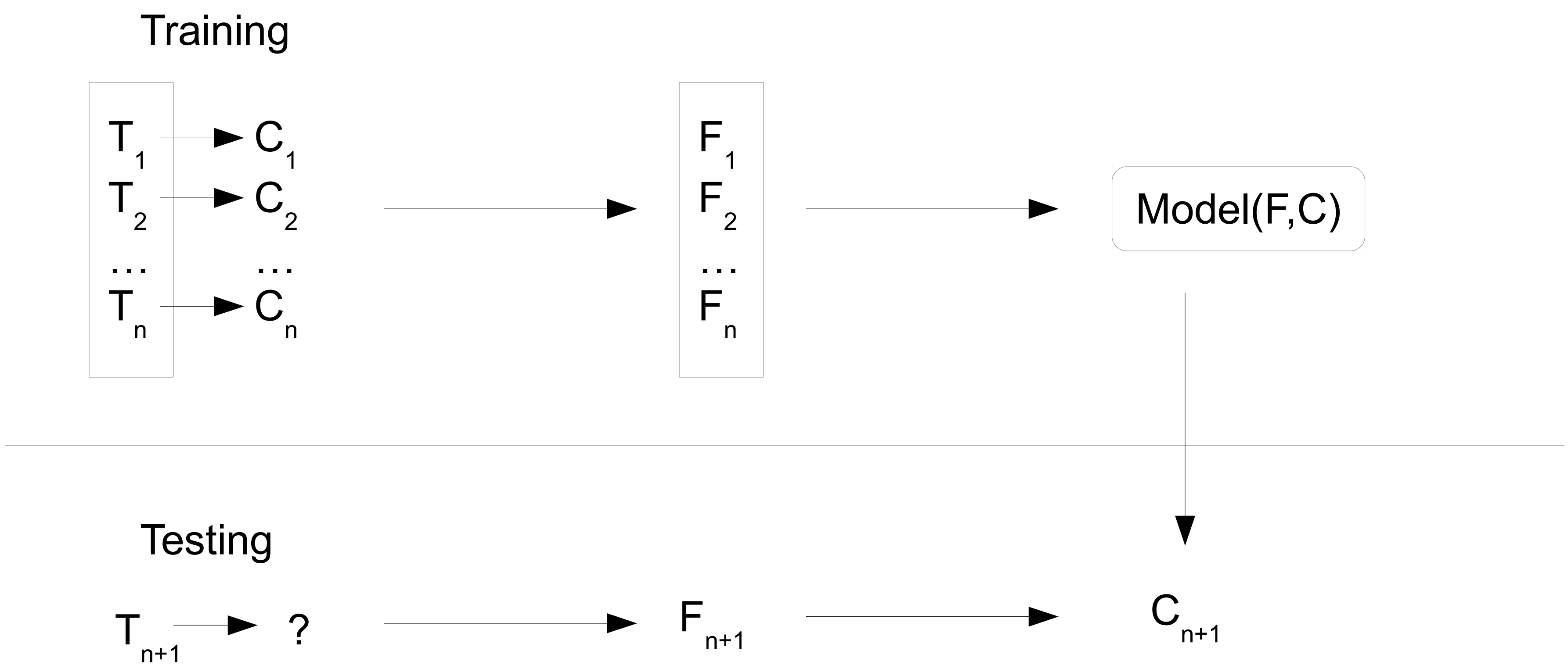
- I$_{[PRP]}$ saw$_{[VBP]}$ the$_{[DT]}$ man$_{[NN]}$ on$_{[IN]}$ the$_{[DT]}$ roof$_{[NN]}$.

- Steven$_{[PER]}$ Paul$_{[PER]}$ Jobs$_{[PER]}$,$_{[O]}$ co-founder$_{[O]}$ of$_{[O]}$ Apple$_{[ORG]}$ Inc$_{[ORG]}$,$_{[O]}$ was$_{[O]}$ born$_{[O]}$ in$_{[O]}$ California$_{[LOC]}$.

Sequence Modeling

- Making a decision based on the
  - Current Observation
    - Word (W_0)
    - Prefix
    - Suffix
    - Lowercased word
    - Capitalization
    - Word shape
  - Surrounding observations
    - W_{+1}
    - W_{-1}
  - Previous decisions
    - T_{-1}
    - T_{-2}

Learning Model



Sequence Modeling

- Greedy inference
  - Starting from the beginning of the sequence
  - Assigning a label to each item using the classifier in that position
  - Using previous decisions as well as the observed data

- Beam inference
  - Keeping the top k labels in each position
  - Extending each sequence in each local way
  - Finding the best k labels for the next position

Hidden Markov Model (HMM)

- Finding the best sequence of tags ($t_1 \dots t_n$) that corresponds to the sequence of observations ($w_1 \dots w_n$)

- 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

$$\hat{t}_1^n = argmax_{t_1^n} P\left(t_1^n | w_1^n\right)$$

Using Bayes Rule

$$\hat{t}_1^n = argmax_{t_1^n} P(t_1^n | w_1^n)$$

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

$$P(t_1^n | w_1^n) = \frac{P(w_1^n | t_1^n) \cdot P(t_1^n)}{P(w_1^n)}$$

$$\hat{t}_1^n = argmax_{t_1^n} \underbrace{P(w_1^n | t_1^n)}_{\text{likelihood}} \cdot \underbrace{P(t_1^n)}_{\text{prior probability}}$$

Using Markov Assumption

$$\hat{t}_1^n = argmax_{t_1^n} P(w_1^n | t_1^n) \cdot P(t_1^n)$$

$$P(w_1^n | t_1^n) \simeq \prod_{i=1}^{n} P(w_i | t_i)$$ (it depends only on its POS tag and independent of other words)

$$P(t_1^n) \simeq \prod_{i=1}^{n} P(t_i | t_{i-1})$$ (it depends only on the previous POS tag, thus, bigram)

$$\hat{t}_1^n = argmax_{t_1^n} \prod_{i=1}^{n} P(w_i | t_i) \cdot P(t_i | t_{i-1})$$

Two Probabilities

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

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

Two Probabilities

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

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

Two Probabilities

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

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

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

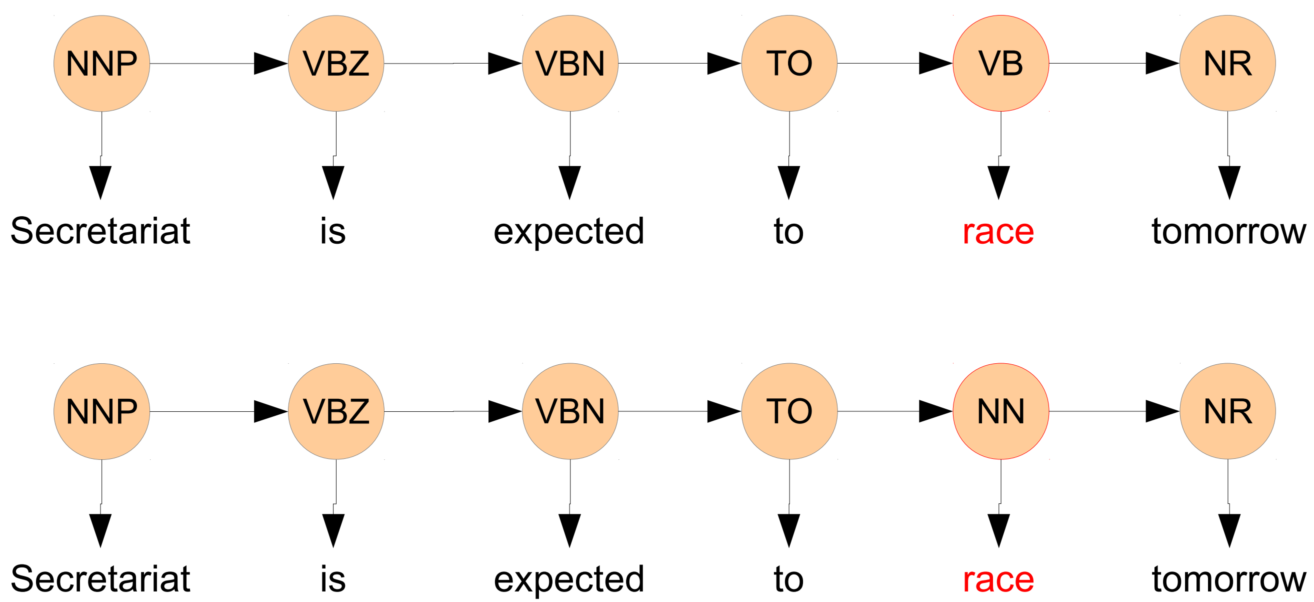
Ambiguity

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$_{[NN]}$.

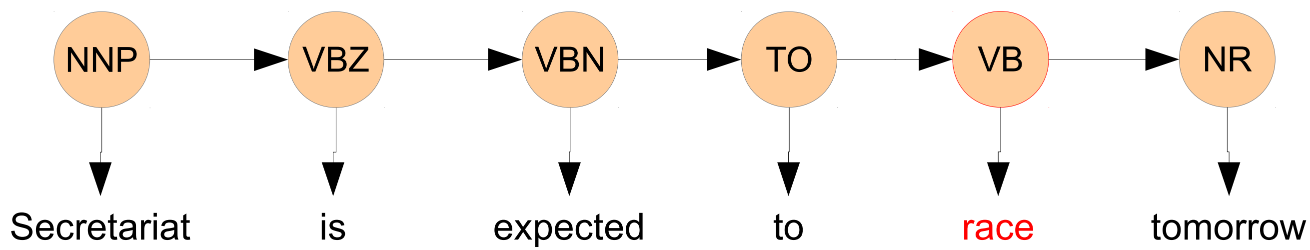
Ambiguity

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



Ambiguity

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



$$P(VB|TO) = 0.83$$

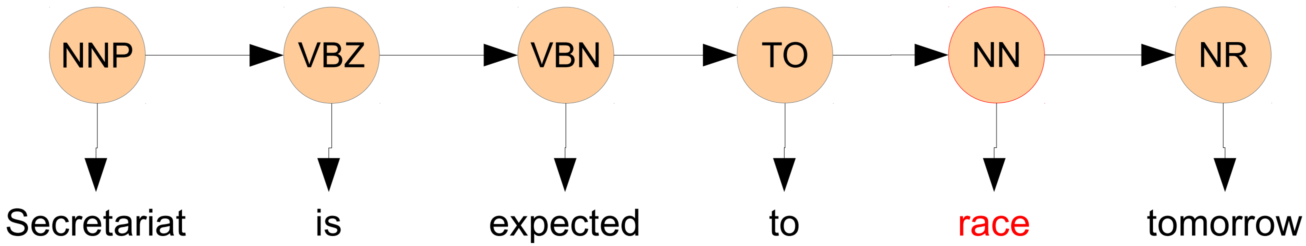
$$P(race|VB) = 0.00012$$

$$P(NR|VB) = 0.0027$$

$$P(VB|TO)P(NR|VB)P(race|VB) = 0.00000027$$

Ambiguity

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



$$P(NN|TO) = 0.00047$$

$$P(race|NN) = 0.00057$$

$$P(NR|NN) = 0.0012$$

$$P(NN|TO)P(NR|NN)P(race|NN) = 0.00000000032$$

Hidden Markov Model (HMM)

- Finite autonom: set of states and a set of transitions between states, according to the input observations
- Weighted finite-state automaton
  - Each arc is associated with a probability
  - The probabilities leaving any arc must sum to one
- Markov chain
  - Special case of weighted autonom
  - Input sequence uniquely determines which states the autonom will go through
  - Useful for assigning probabilities for unambiguous sequences

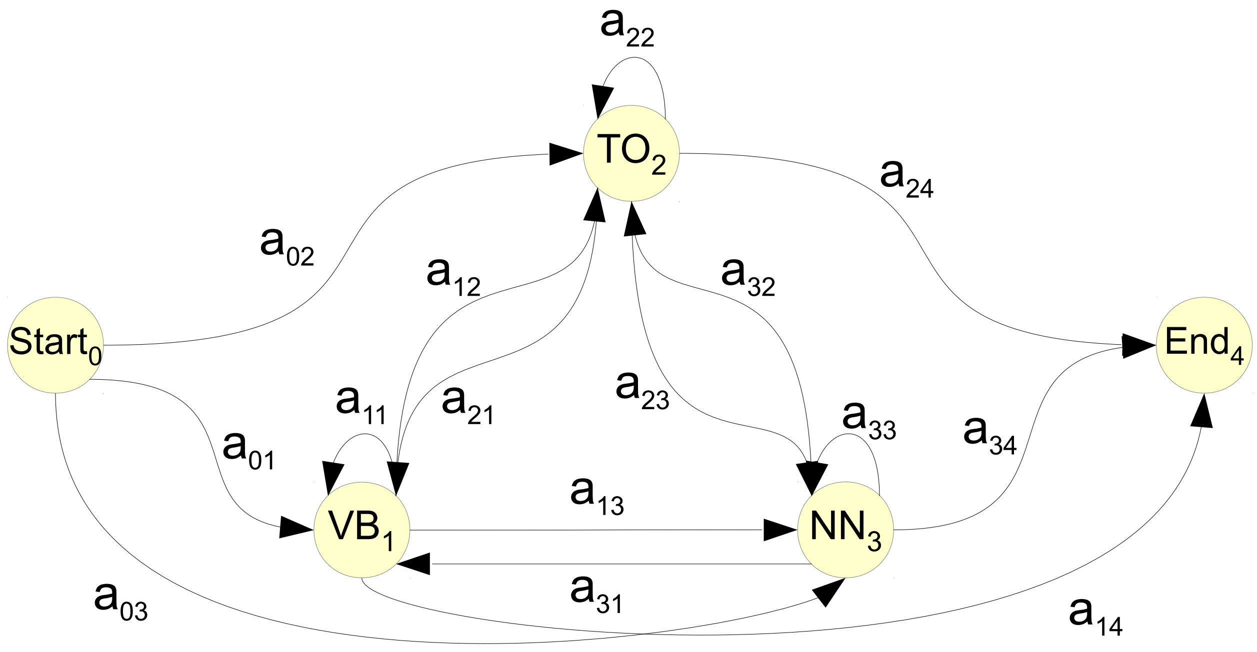
Hidden Markov Model (HMM)

- POS tagging, NER
  - Ambiguous
  - We observe the words, not the POS tags or entity classes

- HMM
  - Observed events: words
  - Hidden events: POS tags, entity classes

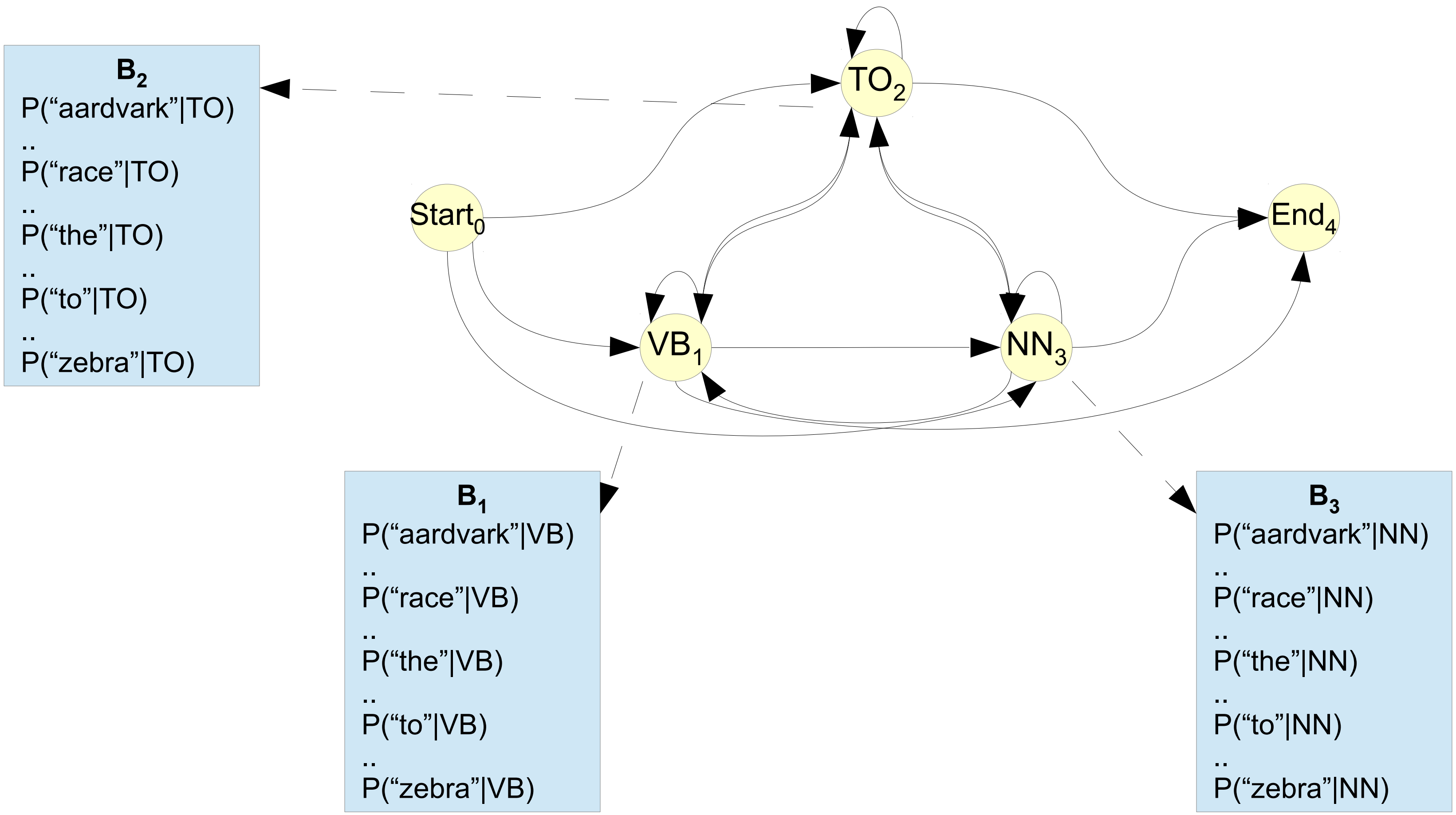
Hidden Markov Model (HMM)

- Transition probabilities: $P(t_i | t_{i-1})$



Hidden Markov Model (HMM)

- Word likelihood probabilities: $P(w_i|t_i)$



The Viterbi Algorithm

- Probability matrix
  - Columns corresponding to inputs (words)
  - Rows corresponding to possible states (POS tags)

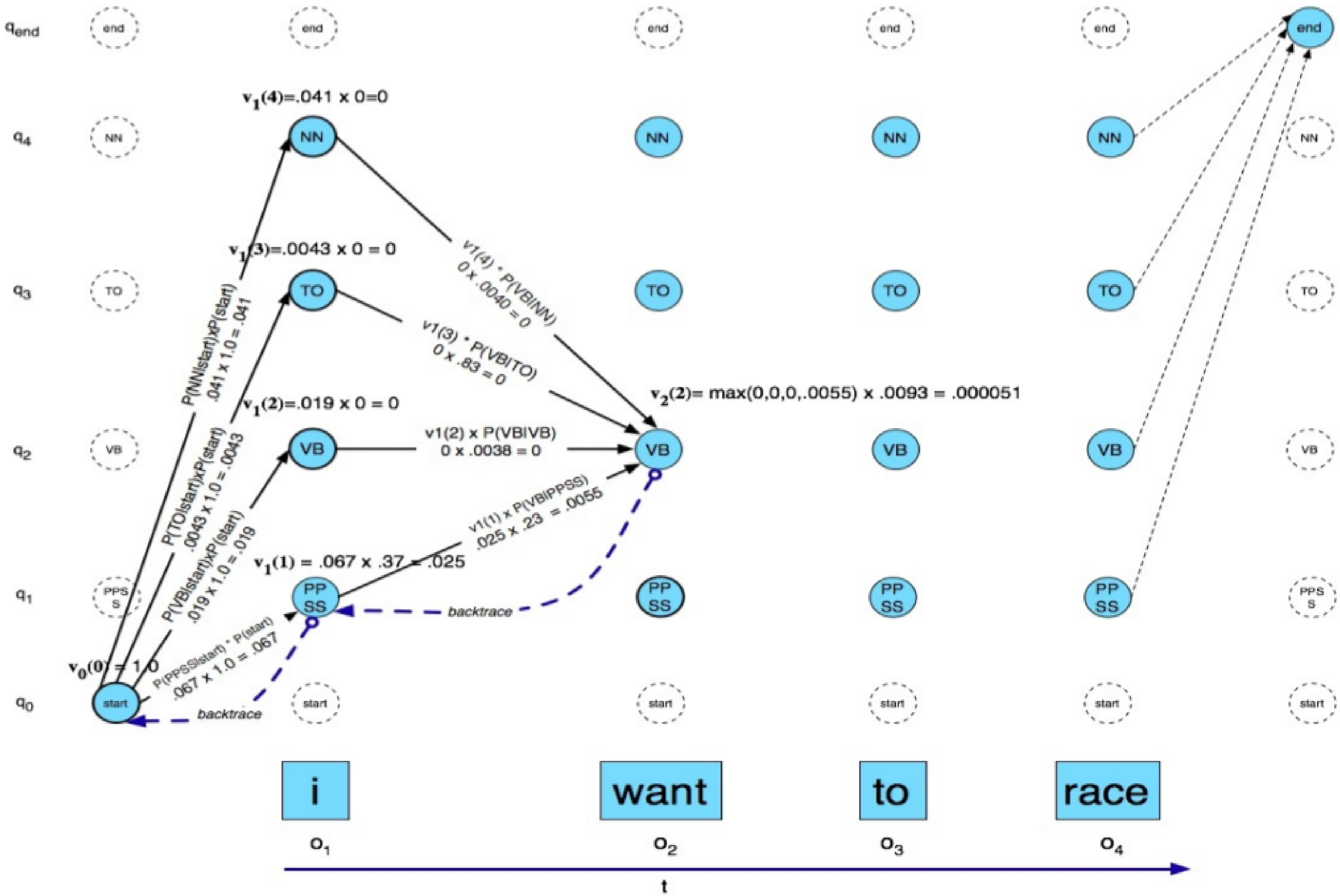
- Move through the matrix in one pass filling the columns left to right using the transition probabilities and observation probabilities

- Storing the max probability path to each cell (not all paths) using dynamic programming

The Viterbi Algorithm

- v_{t-1}: previous Viterbi path probability
  - From the previous time step
- a_{ij}: transition probability
  - From previous state q_i to current state q_j
- $b_j(o_t)$: state observation likelihood
  - Observation symbol o_t given the current state j

The Viterbi Algorithm



Further Reading

- Speech and Language Processing
  - Chapter 5: POS Tagging
  - Chapter 6: MaxEnt & HMM
  - Chapter 22.1: NER