Natural Language Processing
SoSe 2015

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Part-of-speech tagging and named-entity recognition

Dr. Mariana Neves

May 11th, 2015

(based on the slides of Dr. Saeedeh Momtazi)

Outline

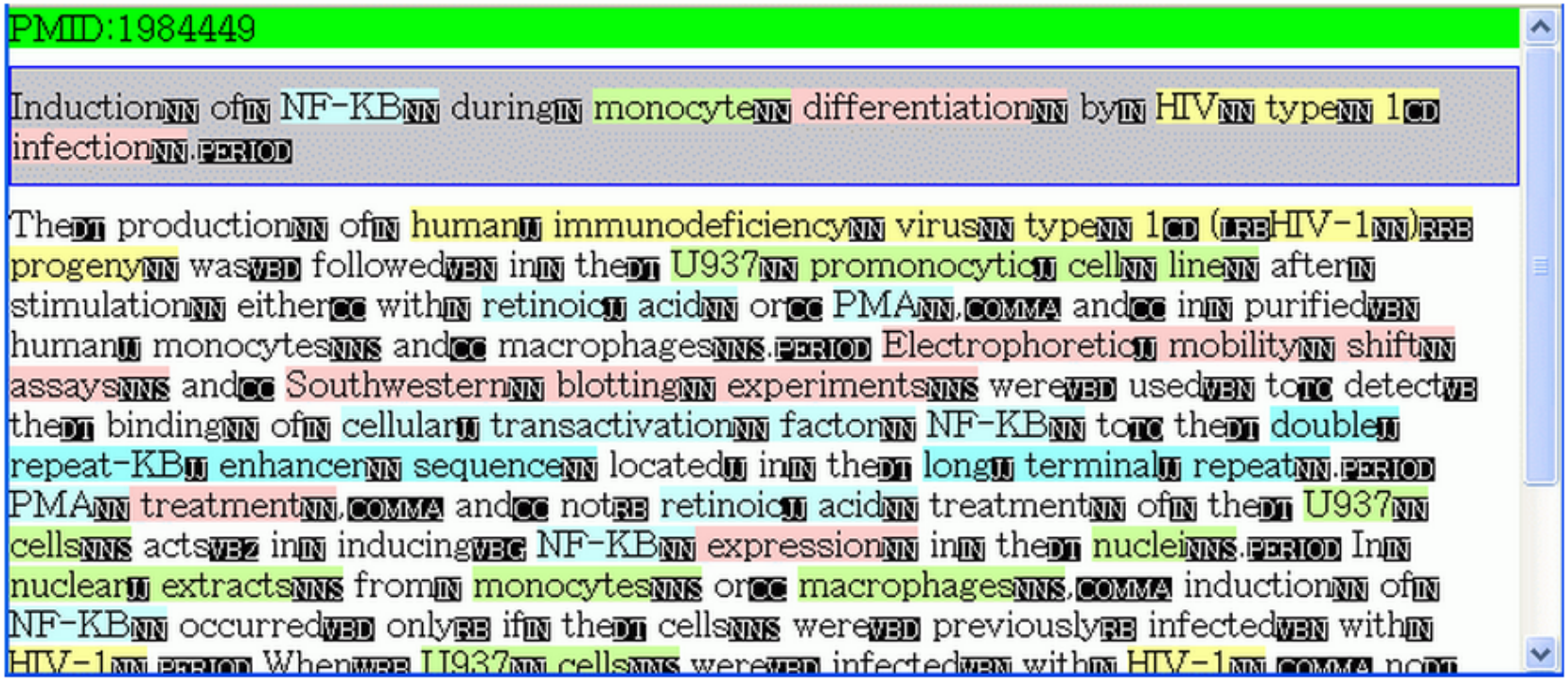
- Part of Speech Tagging
- Named Entity Recognition
- Sequential Modeling

Outline

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

Parts of speech (POS)

- Known as:
  - Parts of speech, lexical categories, word classes, morphological classes, lexical tags



(http://www.nactem.ac.uk/genia/genia-corpus/pos-annotation)

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*
Article/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
- Open
  - unlimited number of words

Open vs. Closed Classes

- Closed

Auxiliary	*can, should, have*
Article/Determiner	*the, some*
Conjunction	*and, or*
Pronoun	*he, my*
Preposition	*to, in*
Particle	*off, up*
Interjection	*Ow, Eh*

Open vs. Closed Classes

- Open

Noun	*book/books, nature, Germany, Sony*
Verb	*eat, wrote*
Adjective	*new, newer, newest*
Adverb	*well, urgently*

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 (named-entity recognition)

```
> 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 many parts of speech tagsets
- 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 (87 tags)
  - Brown corpus
- C5 tagset (61 tags)
- C7 tagset (146 tags!!)
- Penn TreeBank (45 tags)
  - A large annotated corpus of English tagset

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

- 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)

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

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, ...
- It 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

The report said Andreas Lubitz repeatedly set the same plane for an unauthorised descent earlier that day.

Lubitz is suspected of deliberately crashing the Airbus 320, killing all 150 people on board.

He had locked the flight captain out of the cockpit.

Lubitz appears to have practised programming a rapid descent on the outbound leg of the flight - from Duesseldorf to Barcelona on 24 March - the preliminary report by accident investigation agency BEA said.

It added that on several occasions - again with the captain out of the cockpit - the altitude dial was set to 100ft (30m), the lowest possible reading, despite instructions by air traffic control in Bordeaux to set it to 35,000ft and then 21,000ft.

It was also reset on one occasion to 49,000ft, the maximum altitude.

The changes apparently happened over a five-minute period at about 07:30 starting 30 seconds after the captain left the cockpit.

(http://www.bbc.com/news/world-europe-32604552)

Applications

- Finding the main information of a company from its reports
  - Founder, Board members, Headquarters, Profits

Siemens

From Wikipedia, the free encyclopedia

For other uses of "Siemens", see Siemens (disambiguation).

Siemens AG (German pronunciation: [ˈziːmɛns]) is a German multinational conglomerate company headquartered in Berlin and Munich. It is the largest engineering company in Europe. The principal divisions of the company are *Industry*, *Energy*, *Healthcare*, and *Infrastructure & Cities*, which represent the main activities of the company. The company is a prominent maker of medical diagnostics equipment and its medical health-care division, which generates about 12 percent of the company's total sales, is its second-most profitable unit, after the industrial automation division.[2]

Siemens and its subsidiaries employ approximately 343,000 people worldwide and reported global revenue of around €71.9 billion in 2014 according to their annual report.

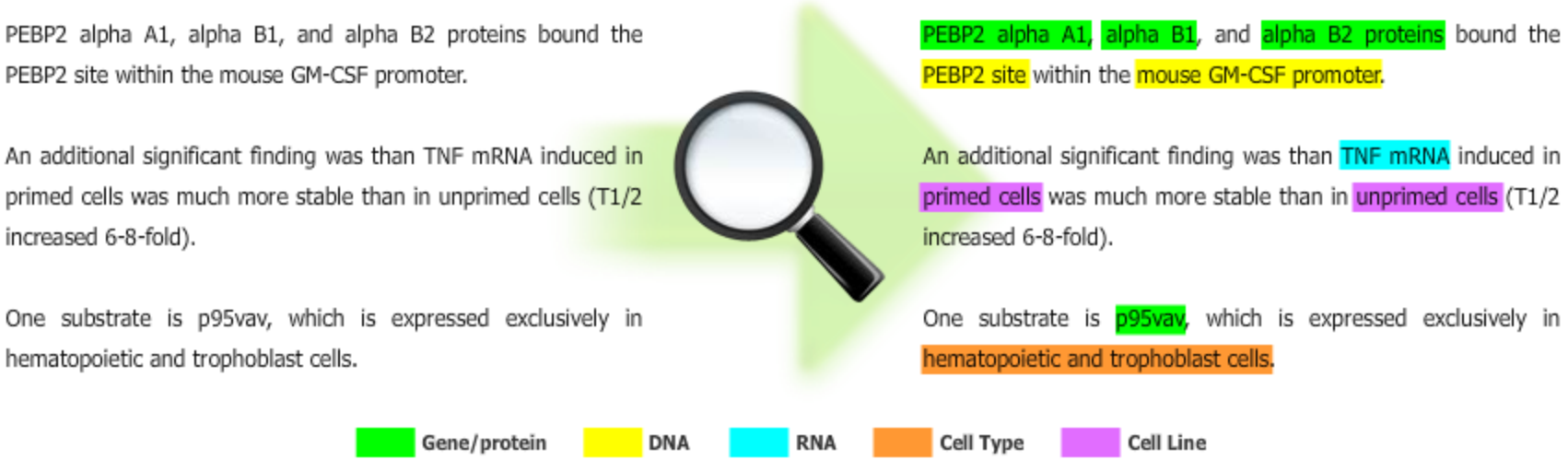
1847 to 1901 [edit]

Siemens & Halske was founded by Werner von Siemens and Johann Georg Halske on 12 October 1847. Based on the telegraph, his invention used a needle to point to the sequence of letters, instead of using Morse code. The company, then called *Telegraphen-Bauanstalt von Siemens & Halske*, opened its first workshop October 12.

(http://en.wikipedia.org/wiki/Siemens)

Applications

- Finding information from biomedical literature
  - Drugs, Genes, Interaction products



(http://bioinformatics.ua.pt/software/gimli/)

Applications

- Finding the target of sentiments
  - Products, Celebrities

An Android User's Review of the iPhone 6: Read this if you are thinking of switching from Android!!!

By Jay on January January 28, 2015

Color Name: Gray | Size Name: 16 GB

So I made the switch from using Android to the iPhone back in October, and I've been using the iPhone 6 for the past few months now and can give a detailed review on what it's like to switch over. Before this switch, I've used the Samsung Galaxy S2 (first smartphone ever!) and also the Nexus 4. Since I'm a tech enthusiast, I'm well versed and have played around with many other Android devices, including all the big names, Galaxy S5, HTC One M8 and M7, One Plus One, and so on. Here are my thoughts:

Things that the iPhone does really well (both hardware and software-wise):
1. Camera. The behind the scene software for digitally capturing an image is definitely the strongest sell for the iPhone. Other than the S5 and Note 4, no smartphone really comes close to having the same kind of image quality (no matter the megapixels) compared to the iPhone. This was one of the reasons for me to switch over since I've started to dabble with photography and wanted a really good camera in my smartphone. (Side note, if you read a lot of tech blogs, there is a notion that in the near future our smartphones won't accurately describe our devices anymore since making a phone call is probably one of the least commonly used features on a smartphone when you look at any average user. Cameras, social media, emails all take a higher usage rate than making a call... really interesting, but anyway, back to the review).

2. Reliability. There have been maybe 2 or 3 times when my phone crashed and would have to be restarted, mostly due to playing some game that was not written very well for the iOS devices. Read more ›

(http://www.amazon.com/Apple-iPhone-Space-Gray-Unlocked/dp/B00NQGP42Y/ref=sr_1_1?s=wireless&ie=UTF8&qid=1431337473&sr=1-1&keywords=iphone)

Named Entity Recognition (NER)

- Finding named entities in a text
- Classifying them to the corresponding classes
- Assigning a unique identifier from a database

- „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]."
- „Steven Paul Jobs [Steve_Jobs], co-founder of Apple Inc [Apple_Inc.], was born in California [California]."

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
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

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 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%

Named entity recognition

- 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] arrived[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, ...

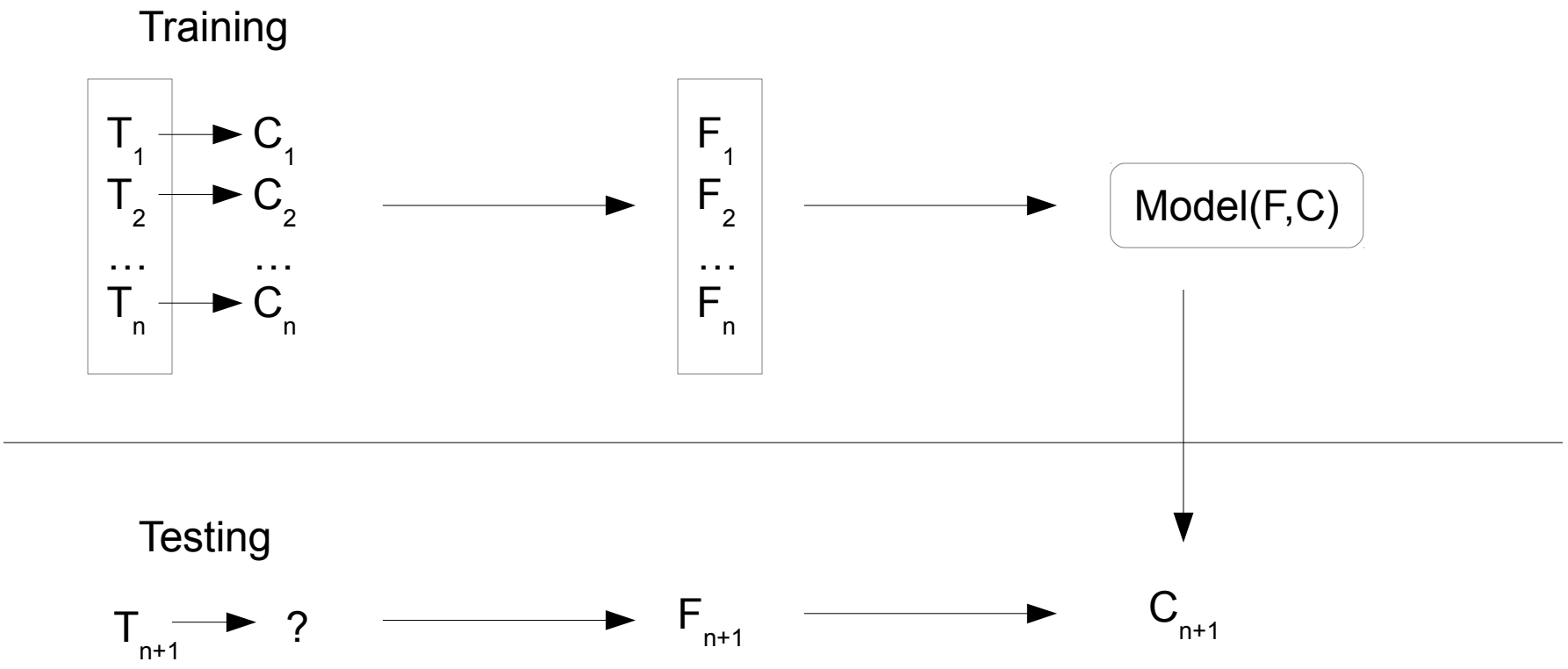
- I[PRP] saw[VBP] the[DT] man[NN] on[IN] the[DT] roof[NN] .

- 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

Sequence Modeling

- 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 ($t_1 \ldots t_n$) 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

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

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[] .

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

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

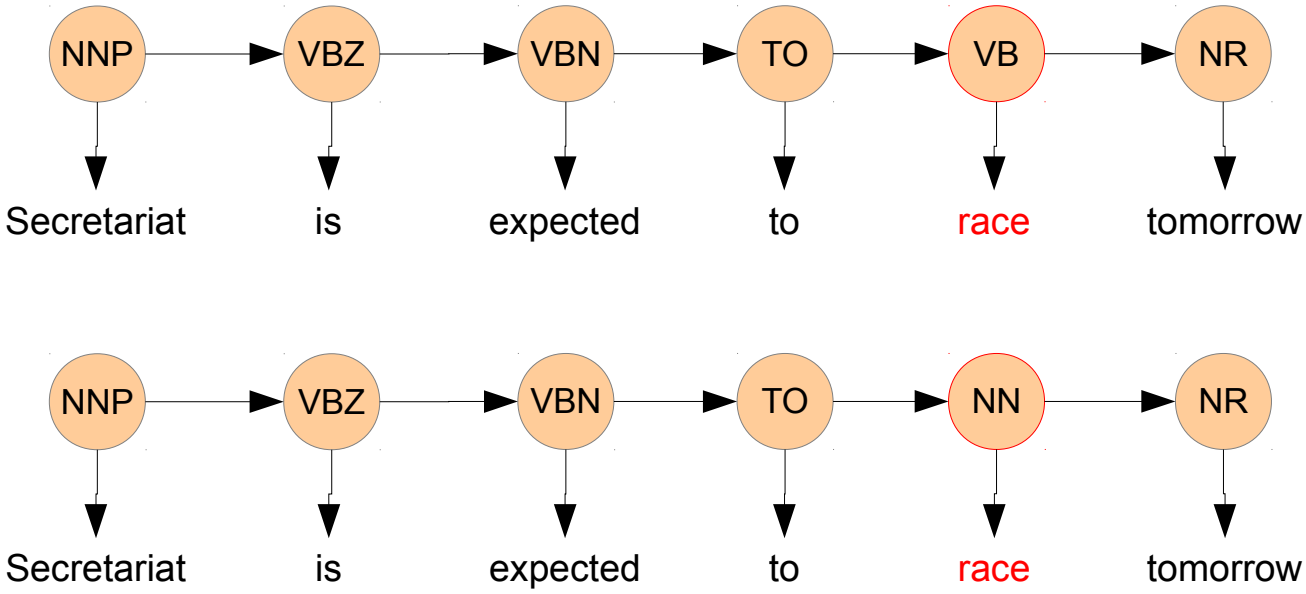
Ambiguity

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] .

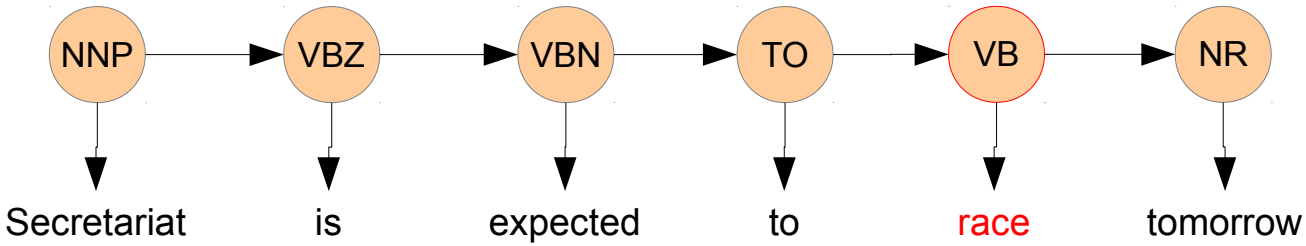
Ambiguity

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] .



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

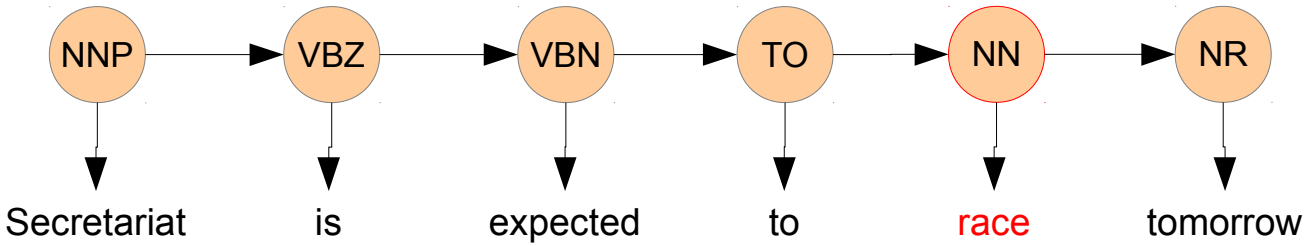
$$P(race|VB) = 0.00012$$

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

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

Ambiguity

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



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

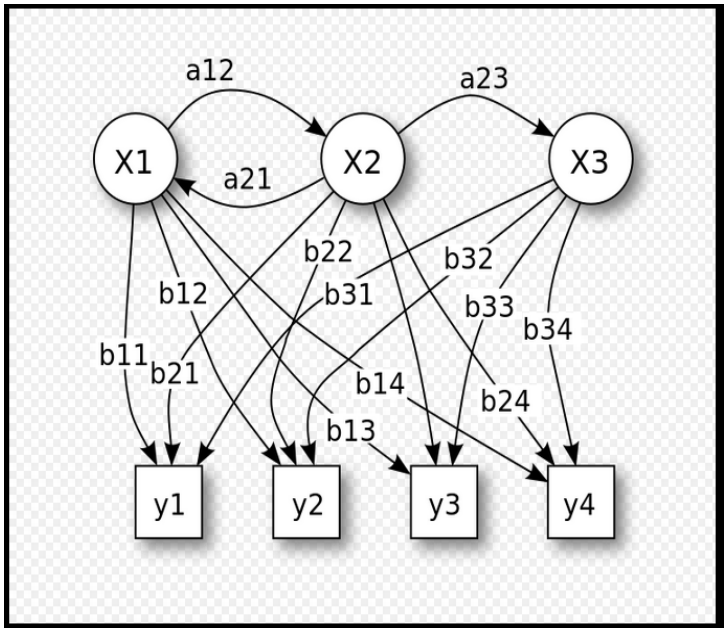
$$P(race|NN) = 0.00057$$

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

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

Hidden Markov Model (HMM)

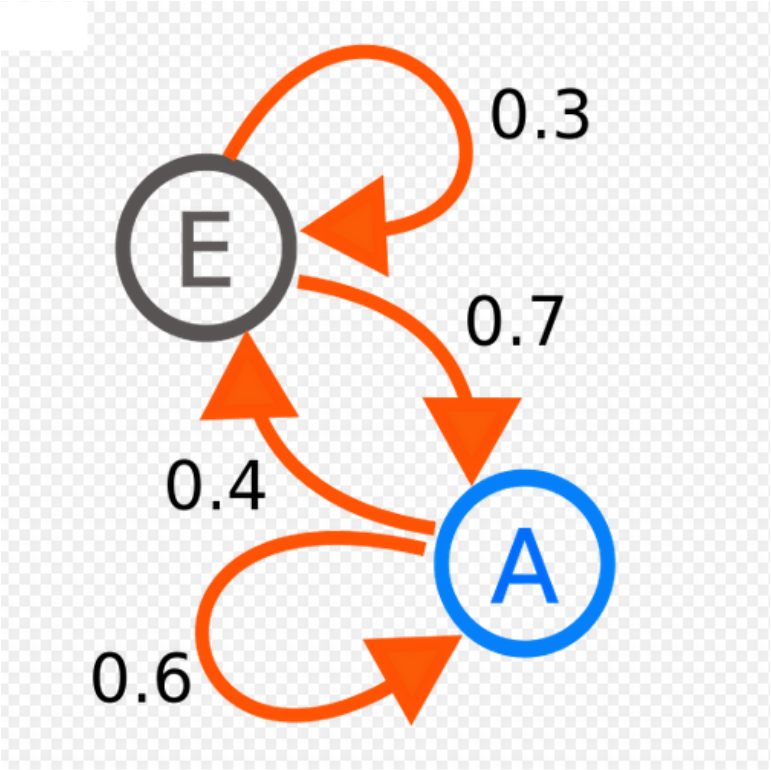
- Finite autonom:
  - set of states
  - set of transitions between states



(http://en.wikipedia.org/wiki/Hidden_Markov_model#/media/File:HiddenMarkovModel.svg)

Hidden Markov Model (HMM)

- Weighted finite-state automaton
  - Each arc is associated with a probability
  - The probabilities leaving any arc must sum to one



(http://en.wikipedia.org/wiki/Markov_chain#/media/File:Markovkate_01.svg)

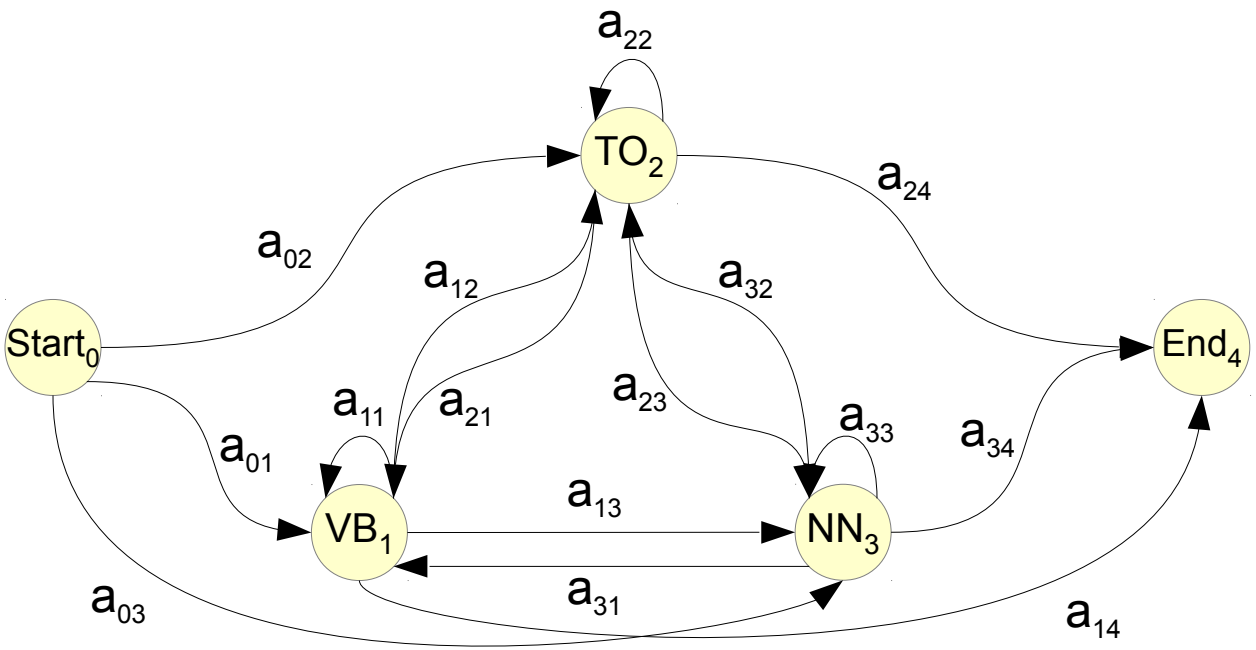
Hidden Markov Model (HMM)

- 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

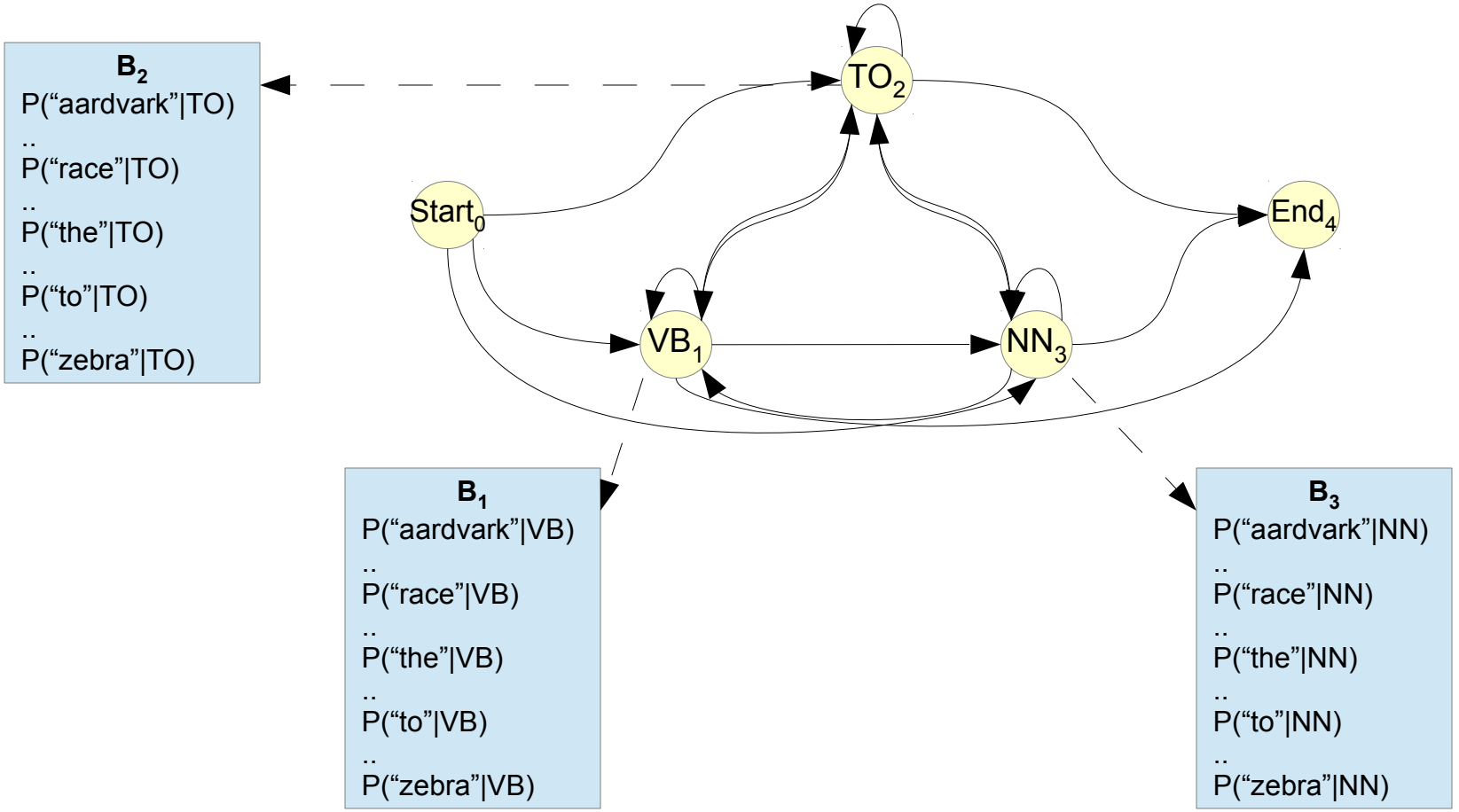
Hidden Markov Model (HMM)

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



Hidden Markov Model (HMM)

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

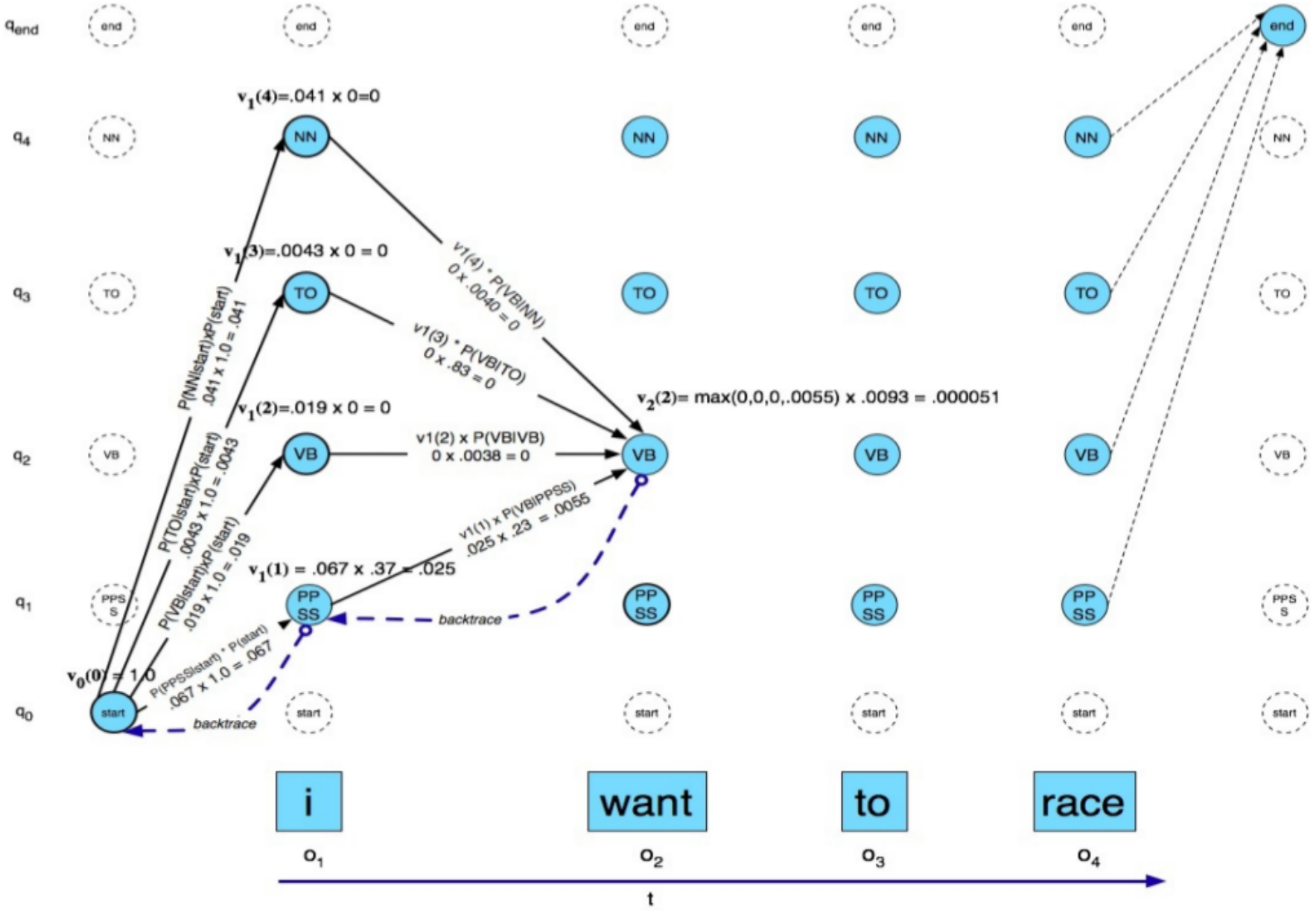


The Viterbi Algorithm

- Decoding algorithm for HMM
- 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
  - Of the observation o_t given the current state j



(Jurafsky and Martin 2009)

Further Reading

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