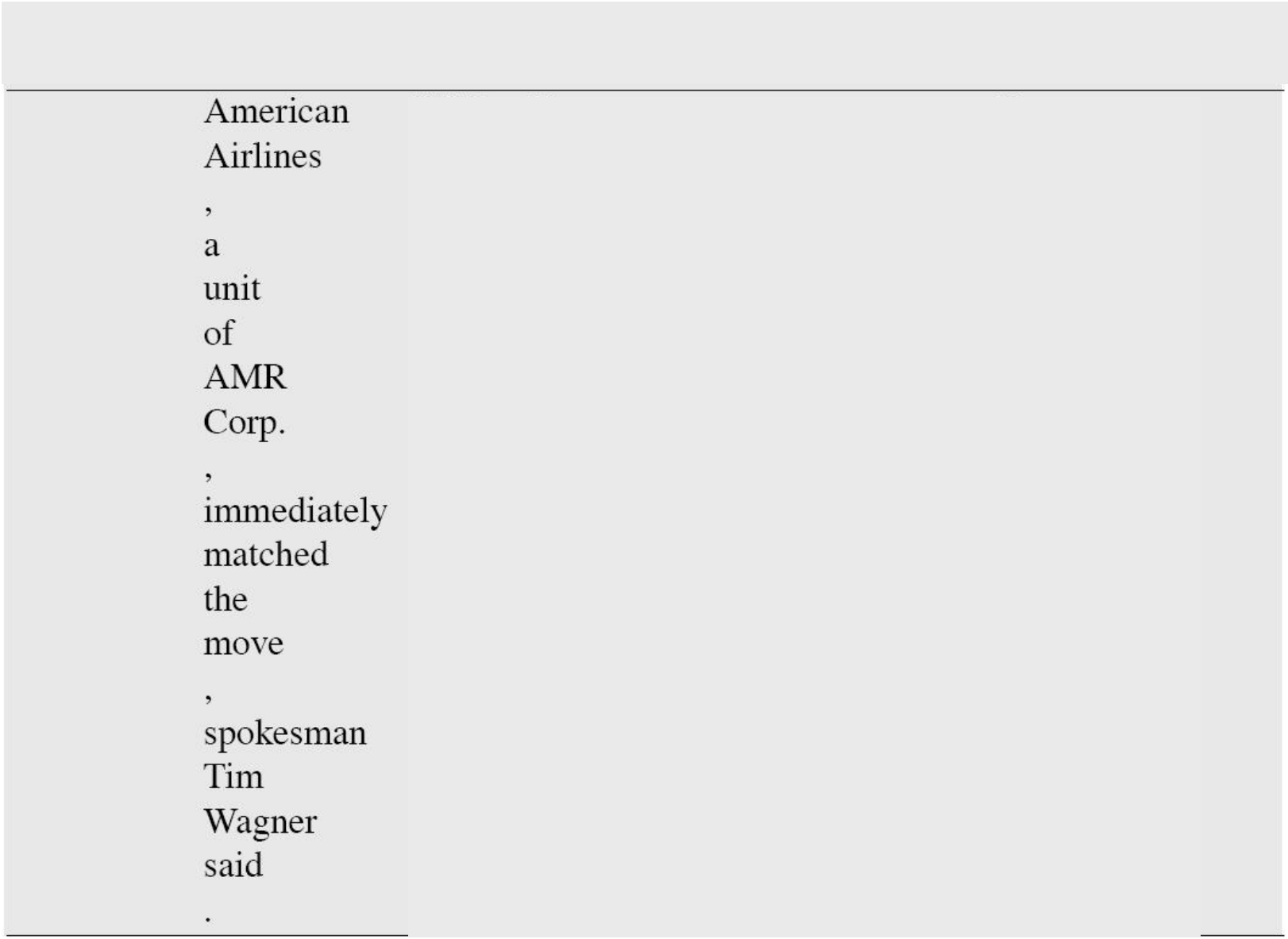
Information Extraction

- **Introduction**
  - Task definition
  - Evaluation
  - IE system architecture
- **Named entity detection --- ML methods**
- **Relation extraction**
  - Manually specified patterns
  - Machine learning approaches

HMMs for NE detection



Figure, copyright J&M 2nd ed

Tag Set for NER

- **IOB tags**
  - B-xxx
    - First (i.e. Beginning) token in a NE of type xxx
  - I-xxx
    - Inside an entity of type xxx
  - O
    - Outside all NEs

HMMs for NE detection

American	B_{ORG}
Airlines	I_{ORG}
,	O
a	O
unit	O
of	O
AMR	B_{ORG}
Corp.	I_{ORG}
,	O
immediately	O
matched	O
the	O
move	O
,	O
spokesman	O
Tim	B_{PER}
Wagner	I_{PER}
said	O
.	O

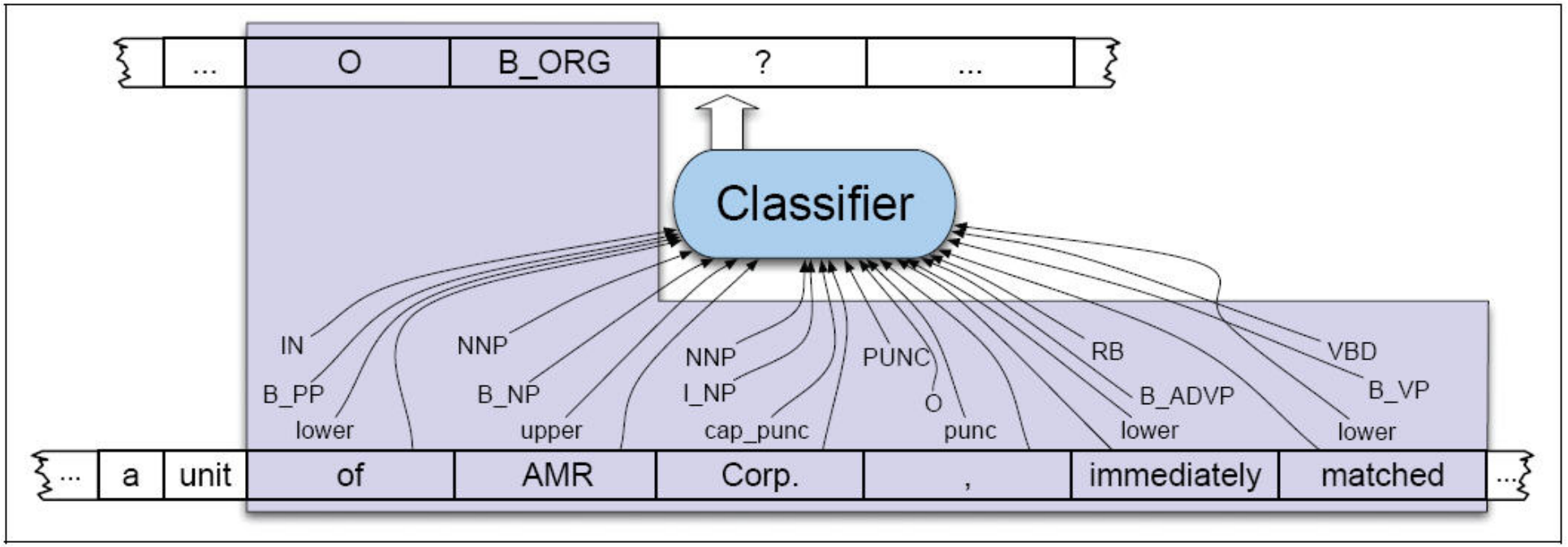
Figure, copyright J&M 2nd ed

HMMs for NE detection

- **Just as for POS tagging except**
  - States are IOB tags

Window-based Classification

- **Fixed-size moving window**
- **Classify the target token as one of IOB**



Figure, copyright J&M 2nd ed

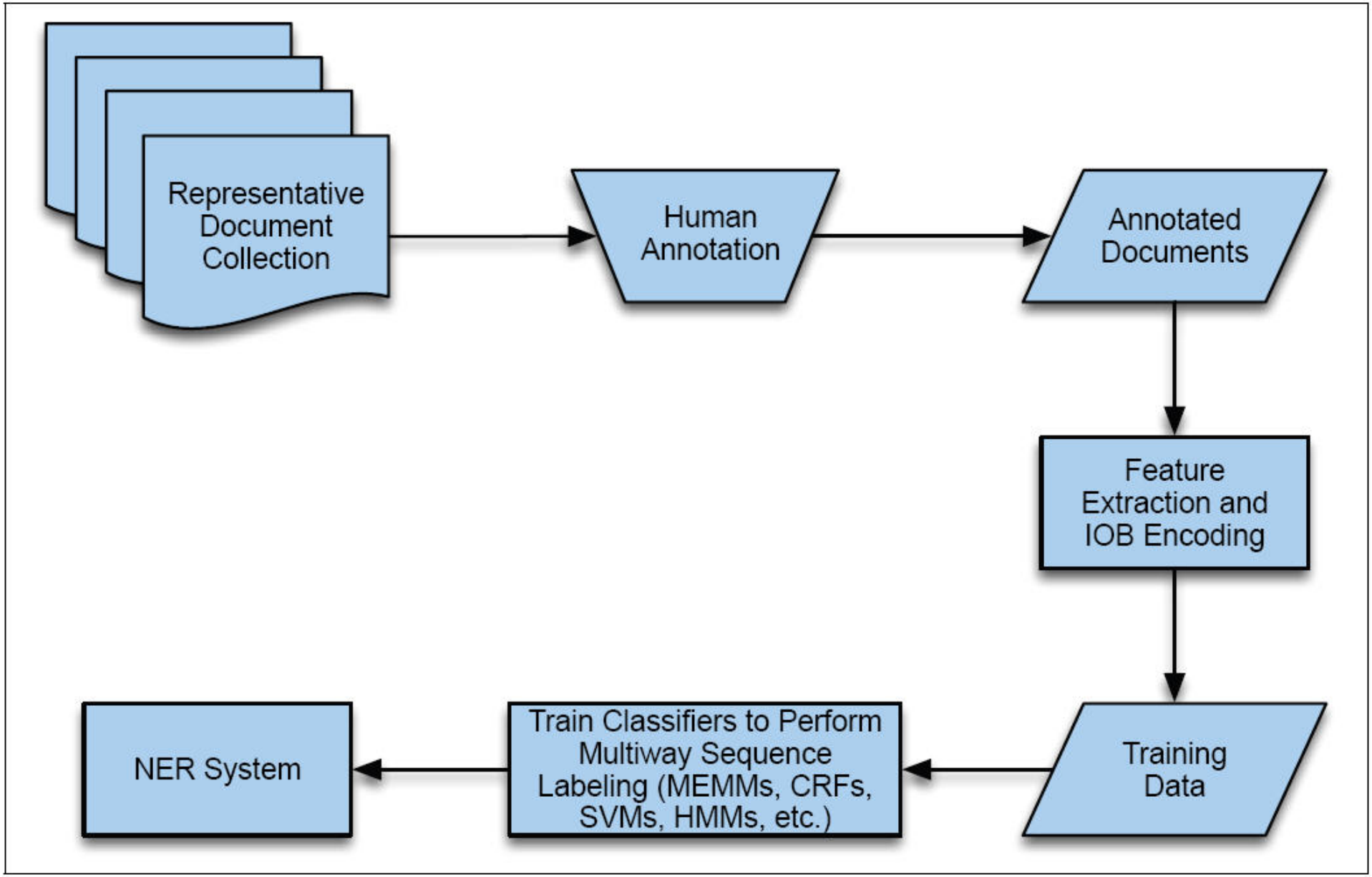
Feature engineering

- **Word features can be especially useful**

Table 3.1 Word features, examples and intuition behind them.[2]

Word Feature	Example Text	Intuition
twoDigitNum	90	Two-digit year
fourDigitNum	1990	Four digit year
containsDigitAndAlpha	A8956-67	Product code
containsDigitAndDash	09-96	Date
containsDigitAndSlash	11/9/89	Date
containsDigitAndComma	23,000.00	Monetary amount
containsDigitAndPeriod	1.00	Monetary amount, percentage
otherNum	456789	Other number
allCaps	BBN	Organization
capPeriod	M.	Person name initial
firstWord	*first word of sentence*	No useful capitalization information
initCap	Sally	Capitalized word
lowerCase	can	Uncapitalized word
other	,	Punctuation marks, all other words

End-to-end process



NE Results Using HMM's

Table 5.1 F-measure Scores. This table illustrates IdentiFinder's performance as compared to the best reported scores for each category.

	Language	Best Rules	HMM
Mixed Case	English (WSJ)	96.4	94.9
Upper Case	English (WSJ)	89	93.6
Speech Form	English (WSJ)	74	90.7
Mixed Case	Spanish	93	90

Information Extraction

- **Introduction**
  - Task definition
  - Evaluation
  - IE system architecture
- **Named entity detection**
- **Relation extraction**
  - Manually specified patterns
  - Machine learning approaches

Person OUT: "Barry Nelson"
OUT position: "president"

VICTIM: "Jesuit priests"
PERP: "the FMLN"

Why bother????
Provide intuition for useful *features* for the machine learning approaches

Acquiring IE patterns

- **Goal**
  - Given a training set of ***annotated*** documents
    - answer keys / gold standard
  - Develop a set of extraction patterns for each slot type.

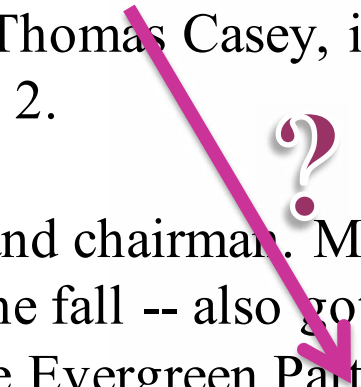
Changes in Management

Evergreen Information said Barry Nelsen, who had a heart-bypass operation last week, resigned as president and chief executive. The board formally accepted the resignation of Thomas Casey, its former chairman, who stepped down effective Feb. 2.

Martin Bell was named president, CEO, and chairman. Mr. Bell -- who has been chief financial officer since the fall -- also got voting control of 970,000 shares held by the Evergreen Partnership, a vehicle for the company's three co-founder

Excluding these shares, Evergreen Informa million shares or exercisable warrants outsta spokeswoman.

The computer products and services concern has cut its staff to fewer than 10 employees from about 35, and has deferred and reduced managers' salaries. In a press release, it said it believes the company is still viable.



In-out-event

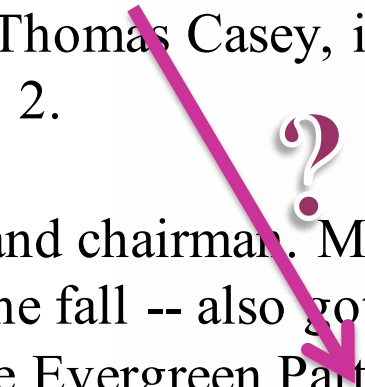
Type:	OUT
Person:	“Barry Nelsen”

Changes in Management

Evergreen Information said Barry Nelsen, who had a heart-bypass operation last week, resigned as president and chief executive. The board formally accepted the resignation of Thomas Casey, its former chairman, who stepped down effective Feb. 2.

Martin Bell was named president, CEO, and chairman. Mr. Bell -- who has been chief financial officer since the fall -- also got voting control of 970,000 shares held by the Evergreen Partnership, a vehicle for the company's three co-founder

Excluding these shares, Evergreen Informa million shares or exercisable warrants outsta spokeswoman.



In-out-event

Type: OUT

Position: PRESIDENT, CHIEF EXECUTIVE

The computer products and services concern has cut its staff to fewer than 10 employees from about 35, and has deferred and reduced managers' salaries. In a press release, it said it believes the company is still viable.

Changes in Management

Evergreen Information said Barry Nelsen, who had a heart-bypass operation last week, resigned as president and chief executive. The board formally accepted the resignation of Thomas Casey, its former chairman, who stepped down effective Feb. 2.

Martin Bell was named president, CEO, and chairman. Mr. Bell -- who has been chief financial officer since the fall -- also got voting control of 970,000 shares held by the Evergreen Partnership, a vehicle for the company's three co-founder

Excluding these shares, Evergreen Informa million shares or exercisable warrants outsta spokeswoman.

The computer products and services concern has cut its staff to fewer than 10 employees from about 35, and has deferred and reduced managers' salaries. In a press release, it said it believes the company is still viable.

?

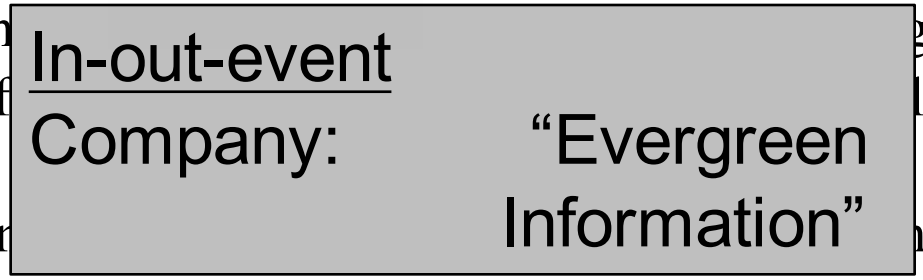
In-out-event

Type:	OUT
Person:	"Barry Nelson"
Position:	"president", "chief executive"

Changes in Management

Evergreen Information said Barry Nelsen, who had a heart-bypass operation last week, resigned as president and chief executive. The board formally accepted the resignation of Thomas Casey, its former chairman, who stepped down effective Feb. 2.

Martin Bell was named president, CEO, and chairman. Mr. Bell -- who has been chief financial officer since the fall -- also got voting con green Partnership, a vehicle f luding Mr. Nelsen.



Excludi as more than two million shares or exercisable warrants outstanding, according to a spokeswoman.

The computer products and services concern has cut its staff to fewer than 10 employees from about 35, and has deferred and reduced managers' salaries. In a press release, it said it believes the company is still viable.

Natural disasters

The twister occurred without warning at approximately 7:15p.m. and destroyed ***two mobile homes***.

Natural disaster	
Type:	TORNADO
Damaged-obj:	“two mobile homes”

Syntactico-semantic patterns

The twister occurred without warning at approximately 7:15p.m. and destroyed ***two mobile homes***.

Pattern:

Trigger: "destroyed"

condition: active voice verb?

Slot: Damaged-Object

Position: direct-object

Also works here…

High winds destroyed ***several office buildings*** in the neighboring town.

Pattern:

Trigger: "destroyed"

condition: active voice verb?

Slot: Damaged-Object

Position: direct-object

And here…uh-oh

The divorce destroyed ***his confidence***.

Pattern:

Trigger: "destroyed"

condition: active voice verb?

Slot: Damaged-Object

Position: direct-object

condition: DO is a physical-object?

Information Extraction

- **Introduction**
  - Task definition
  - Evaluation
  - IE system architecture
- **Named entity detection**
- **Relation extraction**
  - Manually specified patterns
  - Machine learning approaches

Why bother????

Provide intuition for useful *features* for the machine learning approaches

Learning IE patterns from examples

- **Goal**
  - Given a training set of ***annotated*** documents
    - "answer keys" or "gold standard"
    - text spans annotated with slot type
  - Learn extraction patterns for each slot type **using an appropriate machine learning algorithm**.

Learning IE patterns

- **Methods vary with respect to**
  - The **class of pattern** learned (e.g. lexically-based regular expression, syntactic-semantic pattern)
  - **Training corpus** requirements
  - Amount and type of **human feedback** required
  - Degree of **pre-processing** necessary
  - **Other resources**/knowledge bases presumed

Entity and Relation Extraction

- **Learning approaches**
  - Weakly supervised methods
  - Mostly unsupervised methods
  - Sequence-tagging methods
    - MEMM's
  - ILP for joint extraction of entities and relations
    - Opinion extraction

ML Approaches to IE

The twister occurred without warning at approximately 7:15p.m. and destroyed ***two mobile homes***.

Natural disaster

Type:	TORNADO
Damaged-obj:	“two mobile homes”

HMMs for entity detection

American	B_{ORG}
Airlines	I_{ORG}
,	O
a	O
unit	O
of	O
AMR	B_{ORG}
Corp.	I_{ORG}
,	O
immediately	O
matched	O
the	O
move	O
,	O
spokesman	O
Tim	B_{PER}
Wagner	I_{PER}
said	O
.	O

could be Victim, Target, Person-IN, Person-OUT, etc.

Figure, copyright J&M 2nd ed

IOB

$B_{tornado}$	$I_{tornado}$	O	O	O	O	O

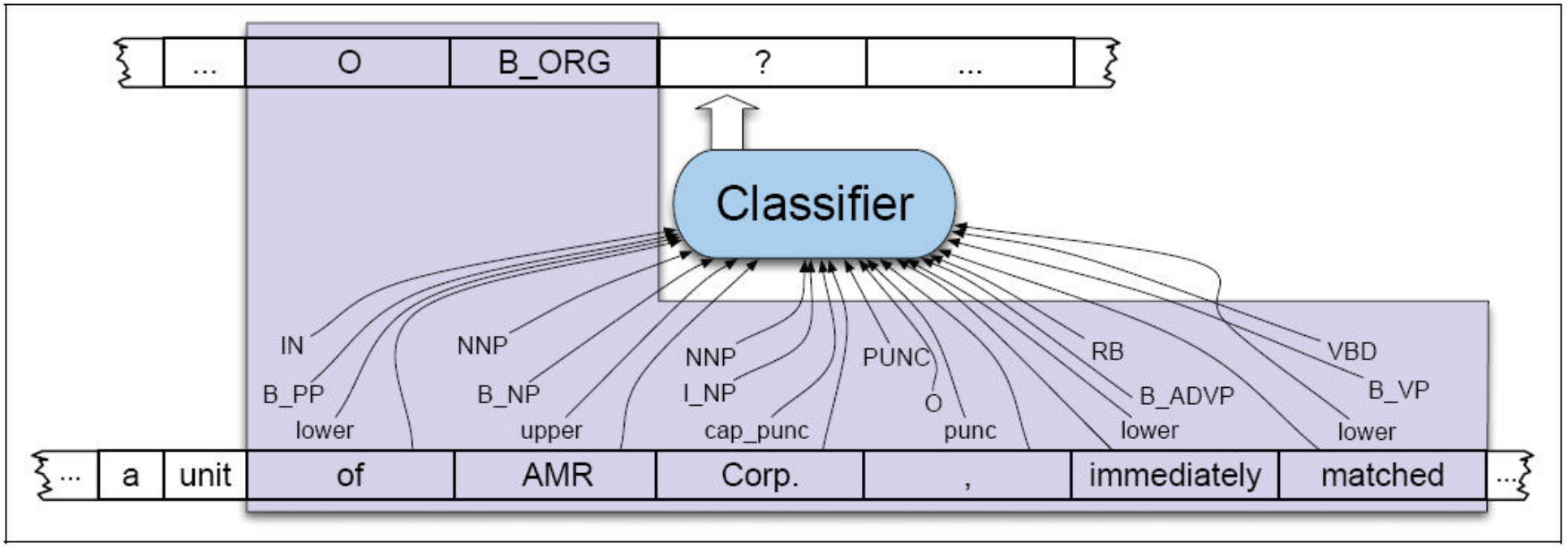
The twister occurred without warning at approximately

B_{time}	I_{time}	O	O	$B_{d\text{-}obj}$	$I_{d\text{-}obj}$	$I_{d\text{-}obj}$	O

7:15 p.m. and destroyed ***two mobile homes*** .

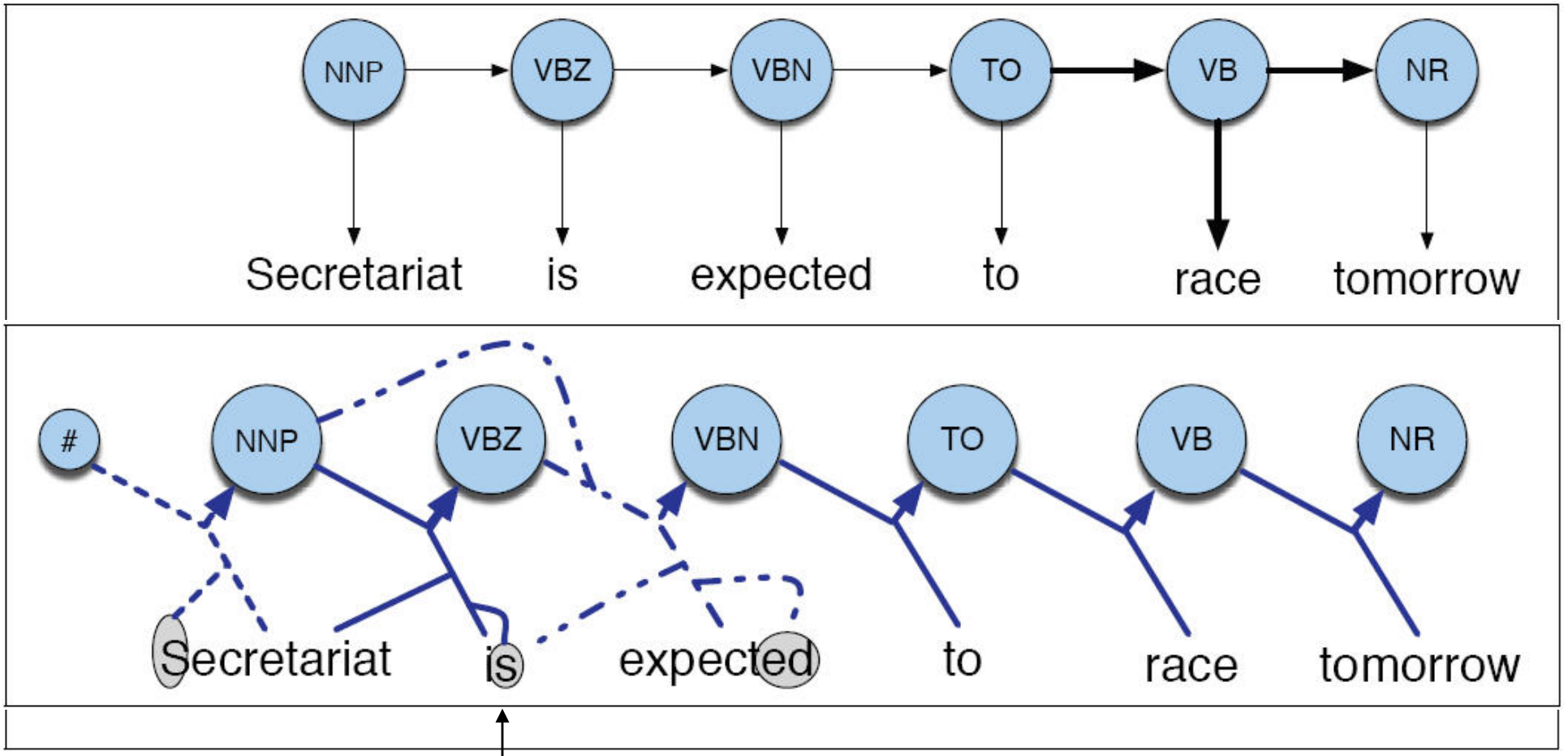
Feature extraction

- **We'd like to be able to include lots of features as in classification-based approaches (e.g. SVMs, dtrees)**



Figure, copyright J&M 2nd ed

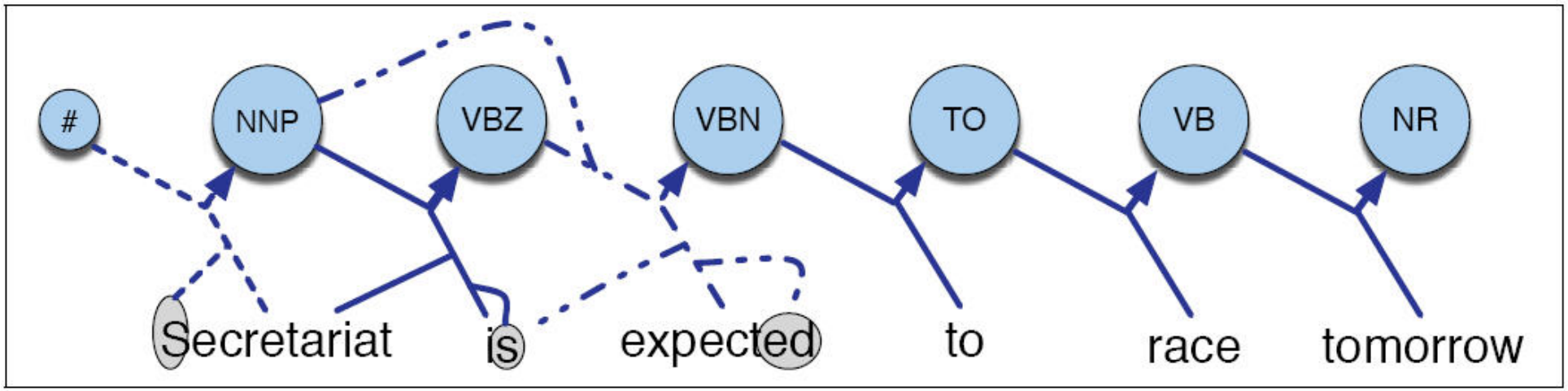
Not possible with HMMs



Figure, copyright J&M 2nd ed

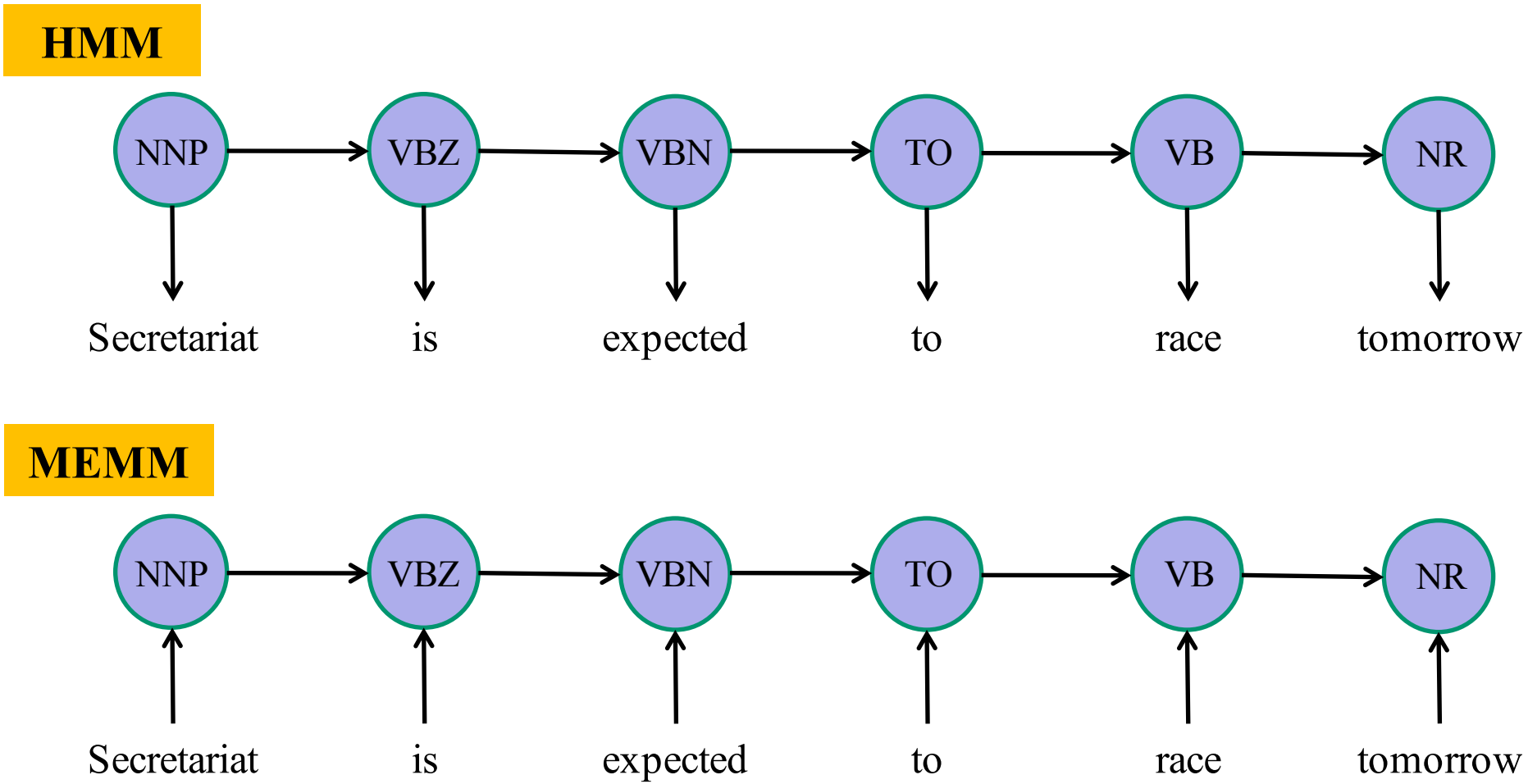
MEMM for p-o-s tagging

- **Can condition on many features of the input**
  - Capitalization
  - Morphology
  - Earlier words
  - Earlier tags

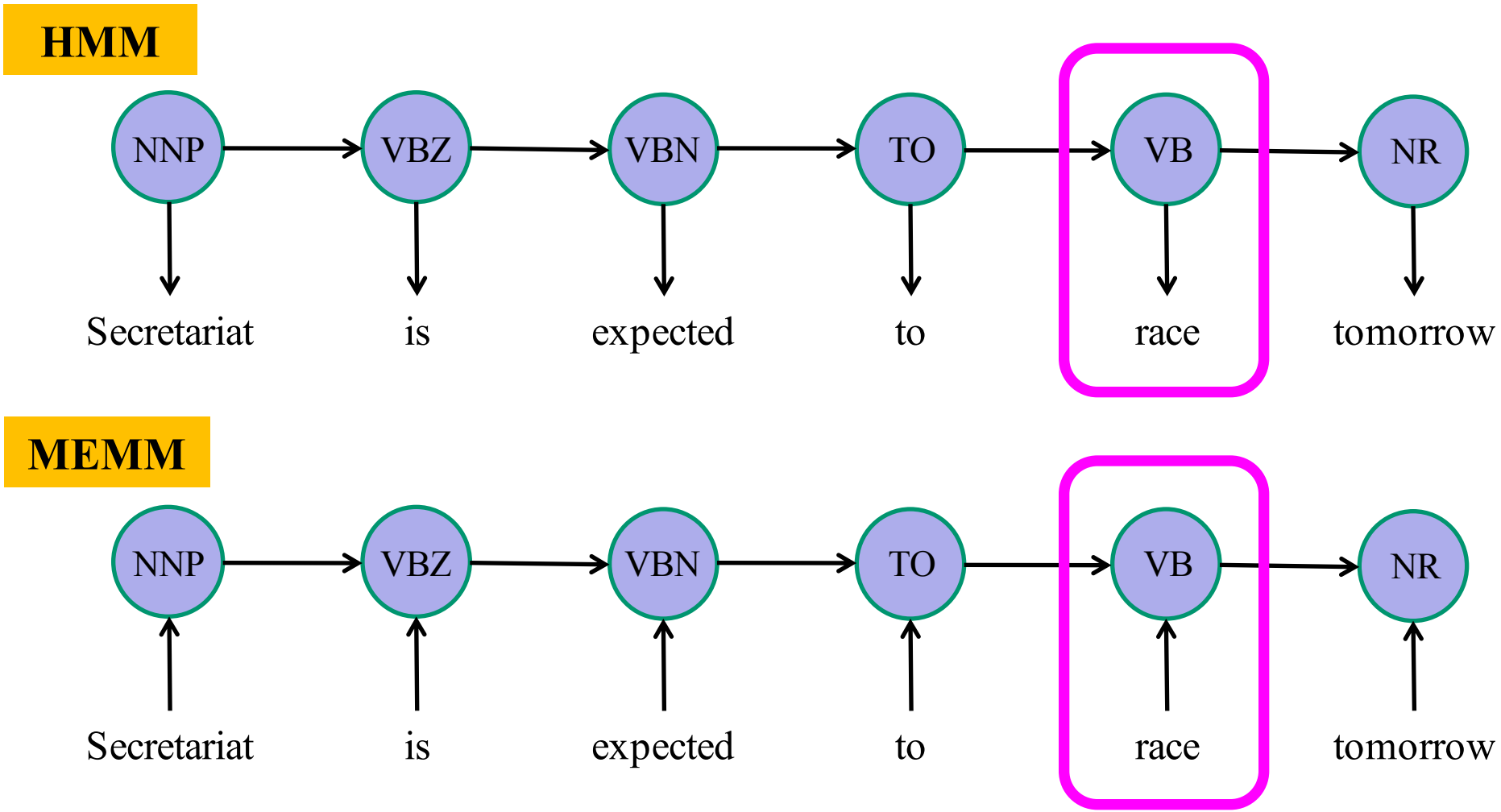


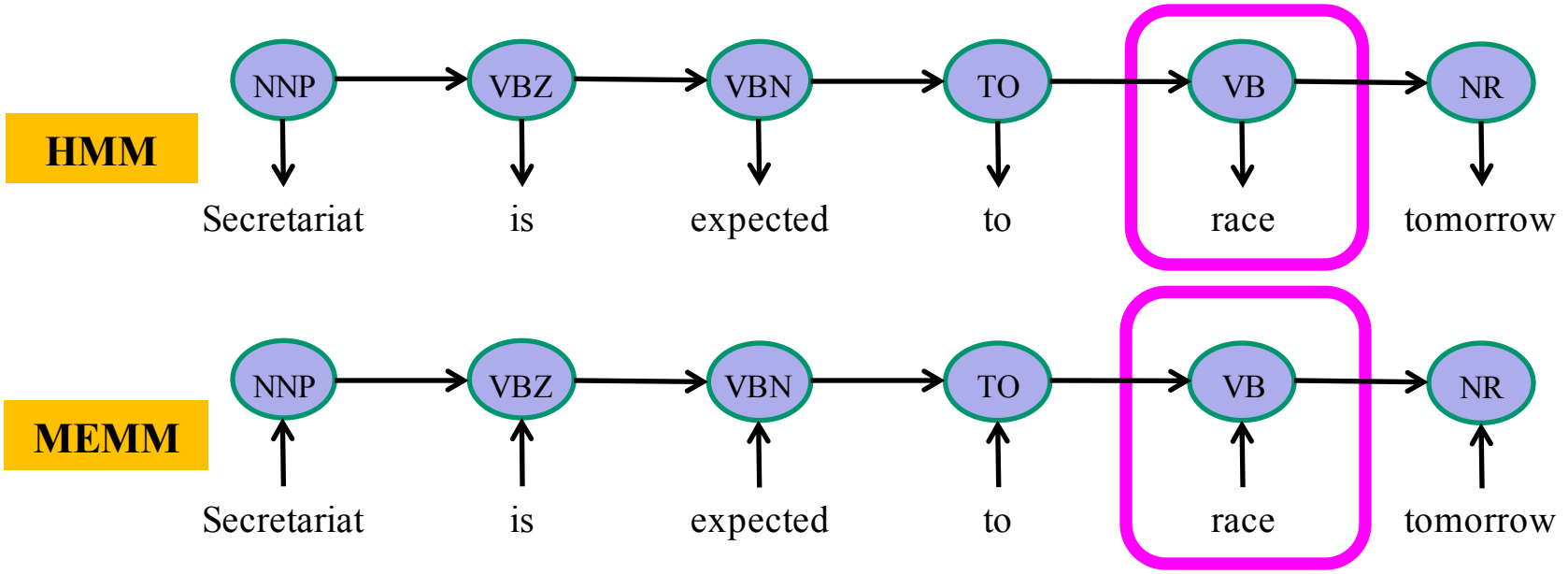
Figure, copyright J&M 2nd ed

HMM vs. MEMM


HMM
NNP
VBZ
VBN
TO
VB
NR
Secretariat
is
expected
to
race
tomorrow
MEMM
NNP
VBZ
VBN
TO
VB
NR
Secretariat
is
expected
to
race
tomorrow

HMM vs. MEMM

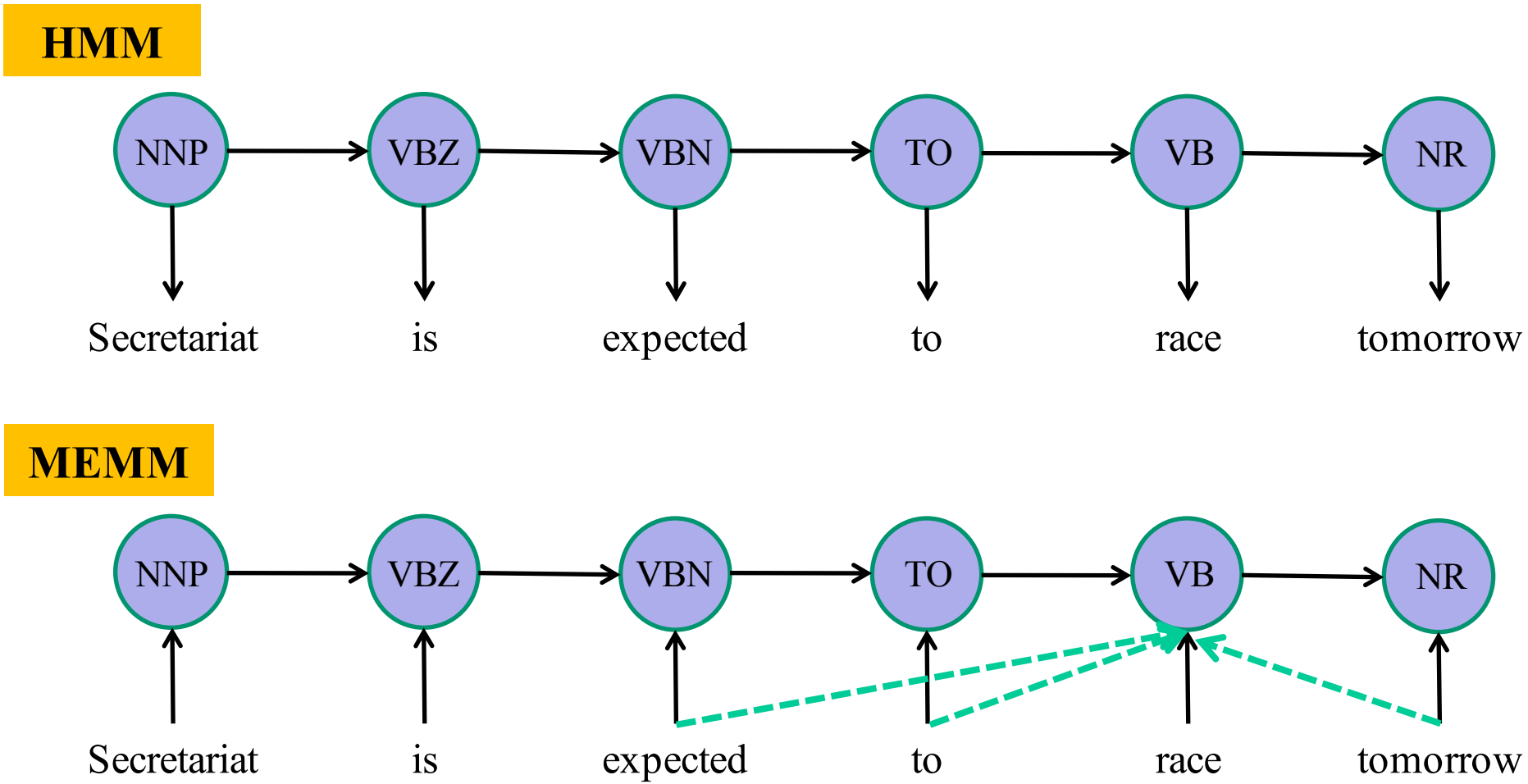

HMM
NNP
VBZ
VBN
TO
VB
NR
Secretariat
is
expected
to
race
tomorrow
MEMM
NNP
VBZ
VBN
TO
VB
NR
Secretariat
is
expected
to
race
tomorrow



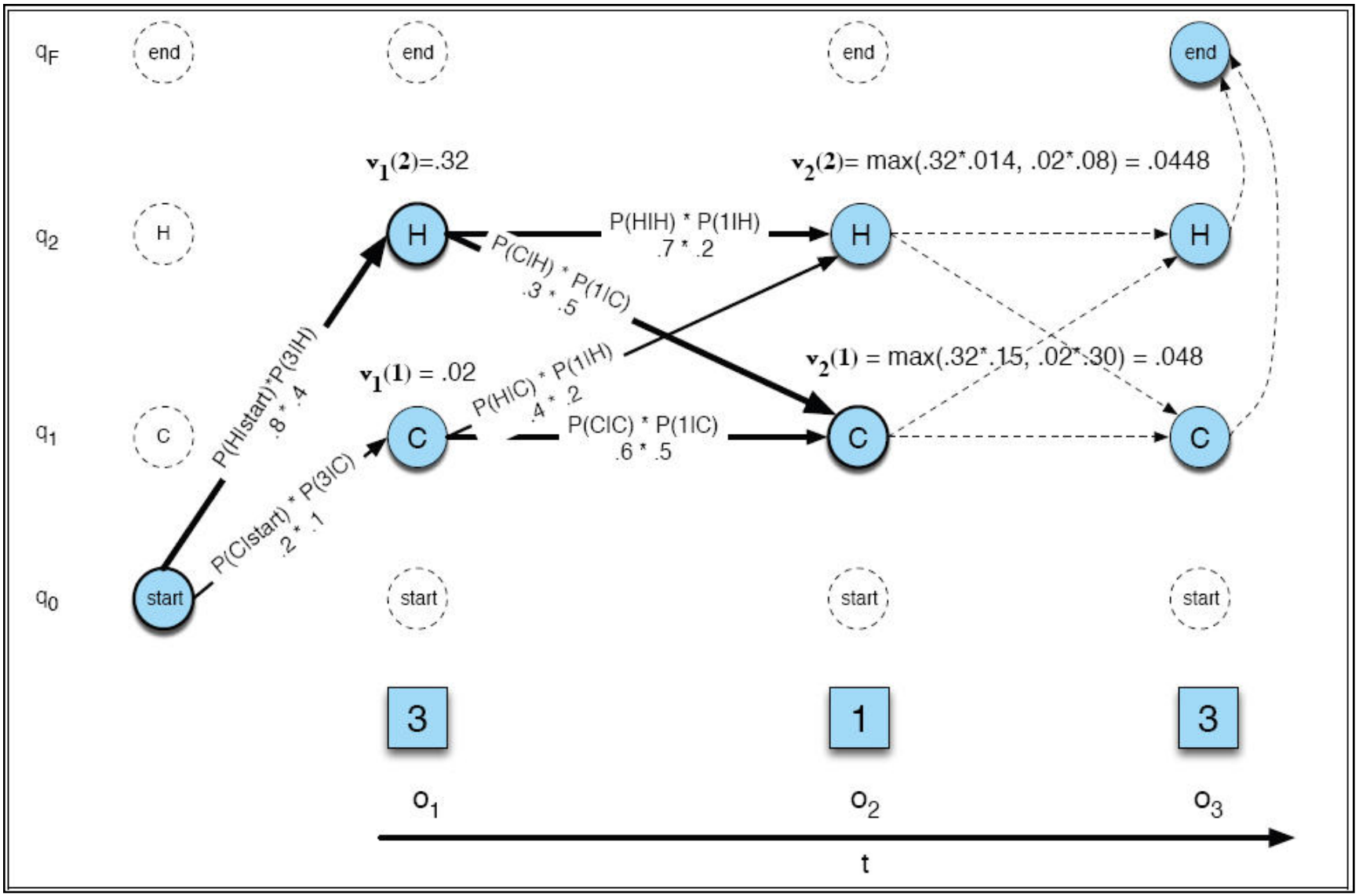
HMM	MEMM
"Generative" models ➔ joint probability **p(words, tags)** ➔"generate" input (in addition to tags) ➔ but we need to predict tags, not words!	"Discriminative" or "Conditional" models ➔ conditional probability **p(tags | words)** ➔"condition" on input ➔ Focusing only on predicting tags
Probability of each slice = transition * emission = p(tag_i | tag_i-1) * p(word_i | tag_i) ➔ Cannot incorporate long distance features	Probability of each slice = p(tag_i | tag_i-1, word_i) or p(tag_i | tag_i-1, all words) ➔ Can incorporate long distance features

via a probabilistic classifier: NB, **MaxEnt**

HMM v.s. MEMM

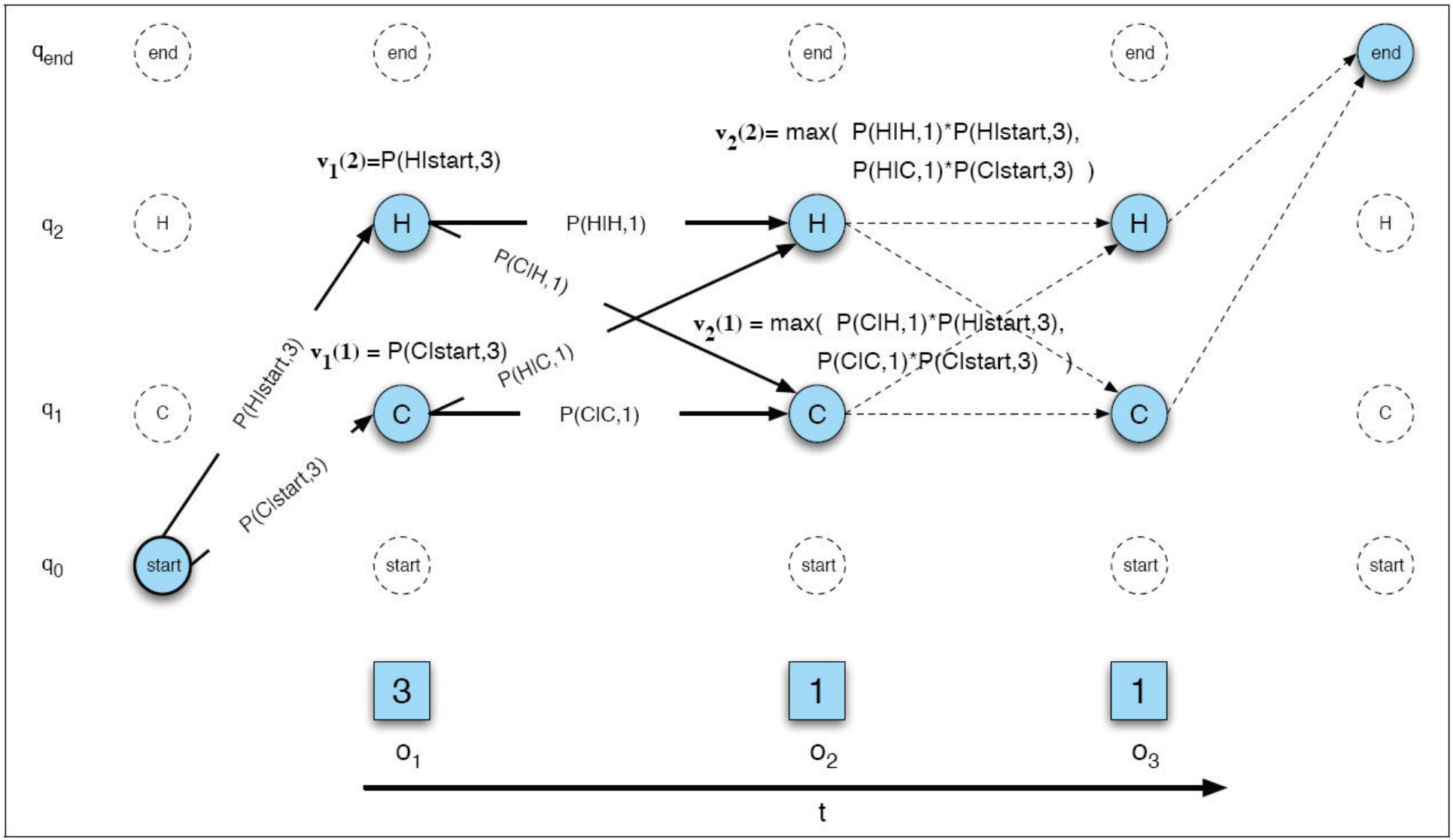

HMM
NNP
VBZ
VBN
TO
VB
NR
Secretariat
is
expected
to
race
tomorrow
MEMM
NNP
VBZ
VBN
TO
VB
NR
Secretariat
is
expected
to
race
tomorrow

Decoding/inference in HMMs



Figure, copyright J&M 2nd ed

Decoding/inference in MEMMs



Figure, copyright J&M 2nd ed

Next…

- **Maximum Entropy classification**