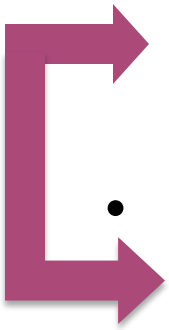
Information extraction

- **Introduction**
  - Task definition
  - Evaluation
  - IE system architecture
- **Acquiring extraction patterns**
  - Manually defined patterns
  - Learning approaches
    - Semi-automatic methods for extraction from unstructured text
    - Fully automatic methods for extraction from structured text
  - Semi-structured text
- **Named entity detection**
- **Sequence-tagging methods for IE**



Information extraction

- **Introduction**
  - Task definition
  - Evaluation
  - IE system architecture
- **Acquiring extraction patterns**
  - Manually defined patterns
  - Learning approaches
    - Semi-automatic methods for extraction from unstructured text
    - Fully automatic methods for extraction from structured text
  - ~~Semi-structured text~~
- **Named entity detection**
- **Sequence-tagging methods for IE**

ML Approaches to Pattern Learning

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

Natural disaster

Type: TORNADO

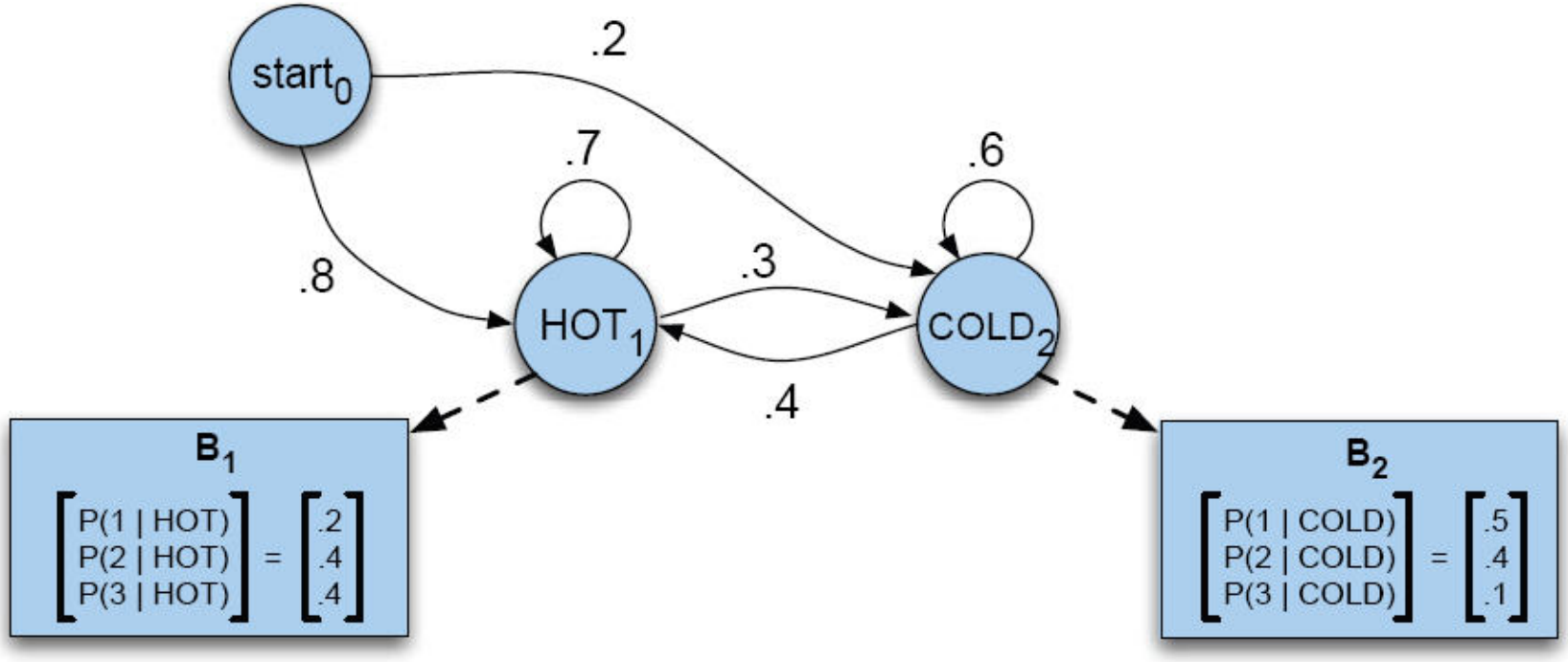
Damaged-obj: “two mobile homes”

Hidden Markov Models

$Q = q_1 q_2 \ldots q_N$	a set of N **states**
$A = a_{11} a_{12} \ldots a_{n1} \ldots a_{nn}$	a **transition probability matrix** A, each a_{ij} representing the probability of moving from state i to state j, s.t. $\sum_{j=1}^{n} a_{ij} = 1 \quad \forall i$
$O = o_1 o_2 \ldots o_T$	a sequence of T **observations**, each one drawn from a vocabulary $V = v_1, v_2, ..., v_V$
$B = b_i(o_t)$	a sequence of **observation likelihoods**, also called **emission probabilities**, each expressing the probability of an observation o_t being generated from a state i
q_0, q_F	a special **start state** and **end (final) state** that are not associated with observations, together with transition probabilities $a_{01} a_{02} \ldots a_{0n}$ out of the start state and $a_{1F} a_{2F} \ldots a_{nF}$ into the end state

Figure, copyright J&M 2nd ed

HMM for weather Prediction



Figure, copyright J&M 2nd ed

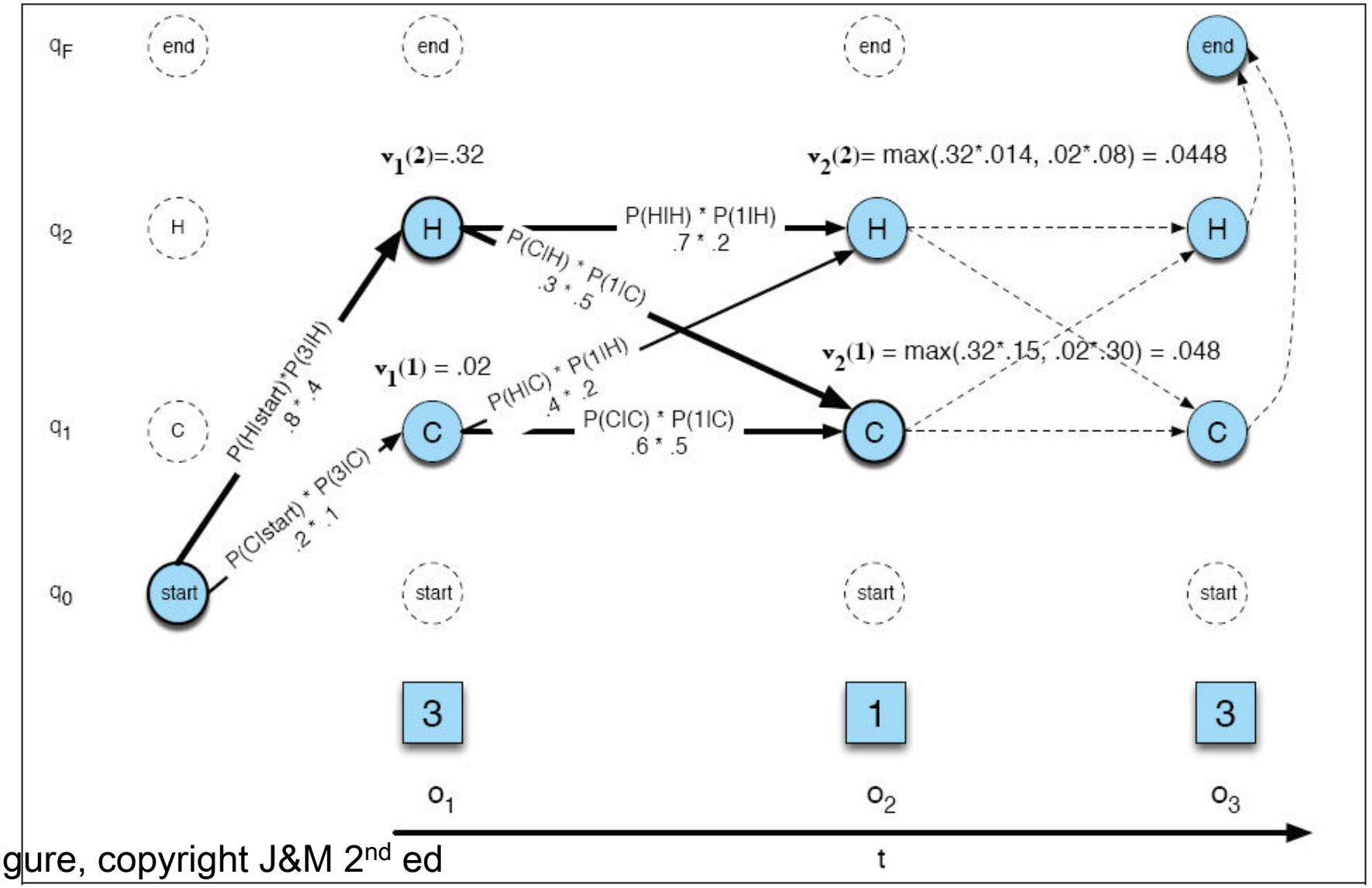
HMMs for entity detection

American	NNP	B_{ORG}
Airlines	NNPS	I_{ORG}
,	PUNC	O
a	DT	O
unit	NN	O
of	IN	O
AMR	NNP	B_{ORG}
Corp.	NNP	I_{ORG}
,	PUNC	O
immediately	RB	O
matched	VBD	O
the	DT	O
move	NN	O
,	PUNC	O
spokesman	NN	O
Tim	NNP	B_{PER}
Wagner	NNP	I_{PER}
said	VBD	O
.	PUNC	O

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

Figure, copyright J&M 2nd ed

Decoding/inference in HMMs



Figure, copyright J&M 2nd ed

Classification approach???

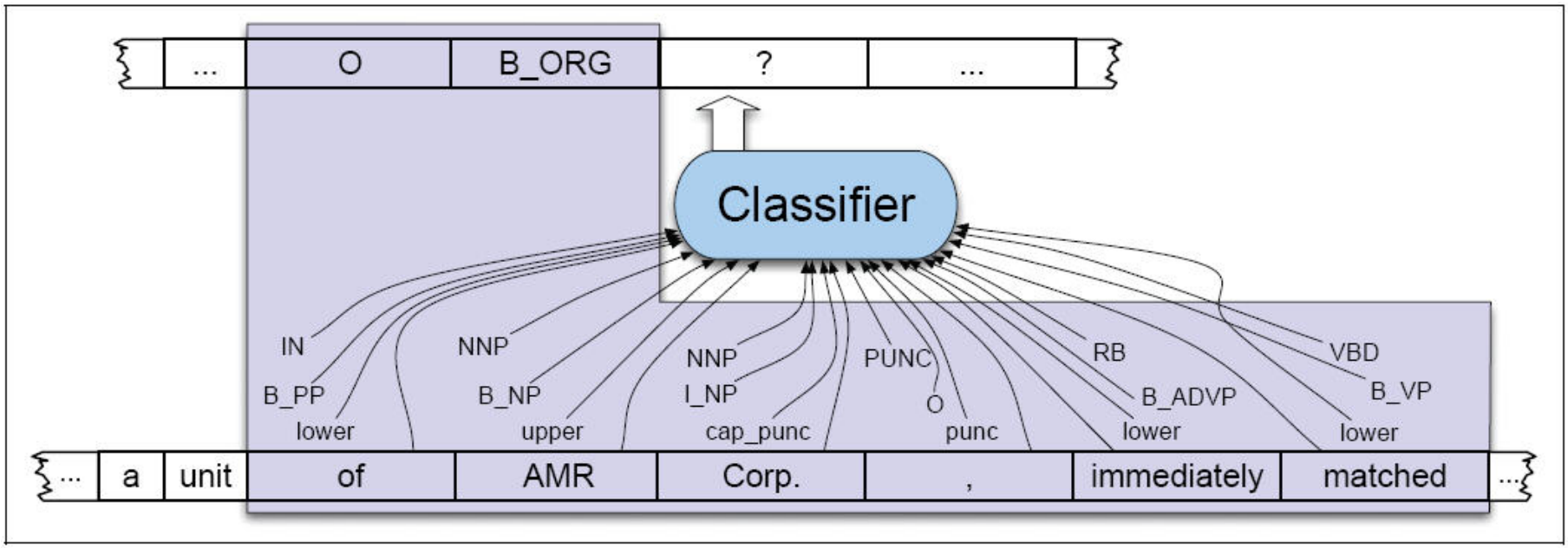
Features				Label
American	NNP	B_{NP}	cap	B_{ORG}
Airlines	NNPS	I_{NP}	cap	I_{ORG}
,	PUNC	O	punc	O
a	DT	B_{NP}	lower	O
unit	NN	I_{NP}	lower	O
of	IN	B_{PP}	lower	O
AMR	NNP	B_{NP}	upper	B_{ORG}
Corp.	NNP	I_{NP}	cap_punc	I_{ORG}
,	PUNC	O	punc	O
immediately	RB	B_{ADVP}	lower	O
matched	VBD	B_{VP}	lower	O
the	DT	B_{NP}	lower	O
move	NN	I_{NP}	lower	O
,	PUNC	O	punc	O
spokesman	NN	B_{NP}	lower	O
Tim	NNP	I_{NP}	cap	B_{PER}
Wagner	NNP	I_{NP}	cap	I_{PER}
said	VBD	B_{VP}	lower	O
.	PUNC	O	punc	O

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

Figure, copyright J&M 2nd ed

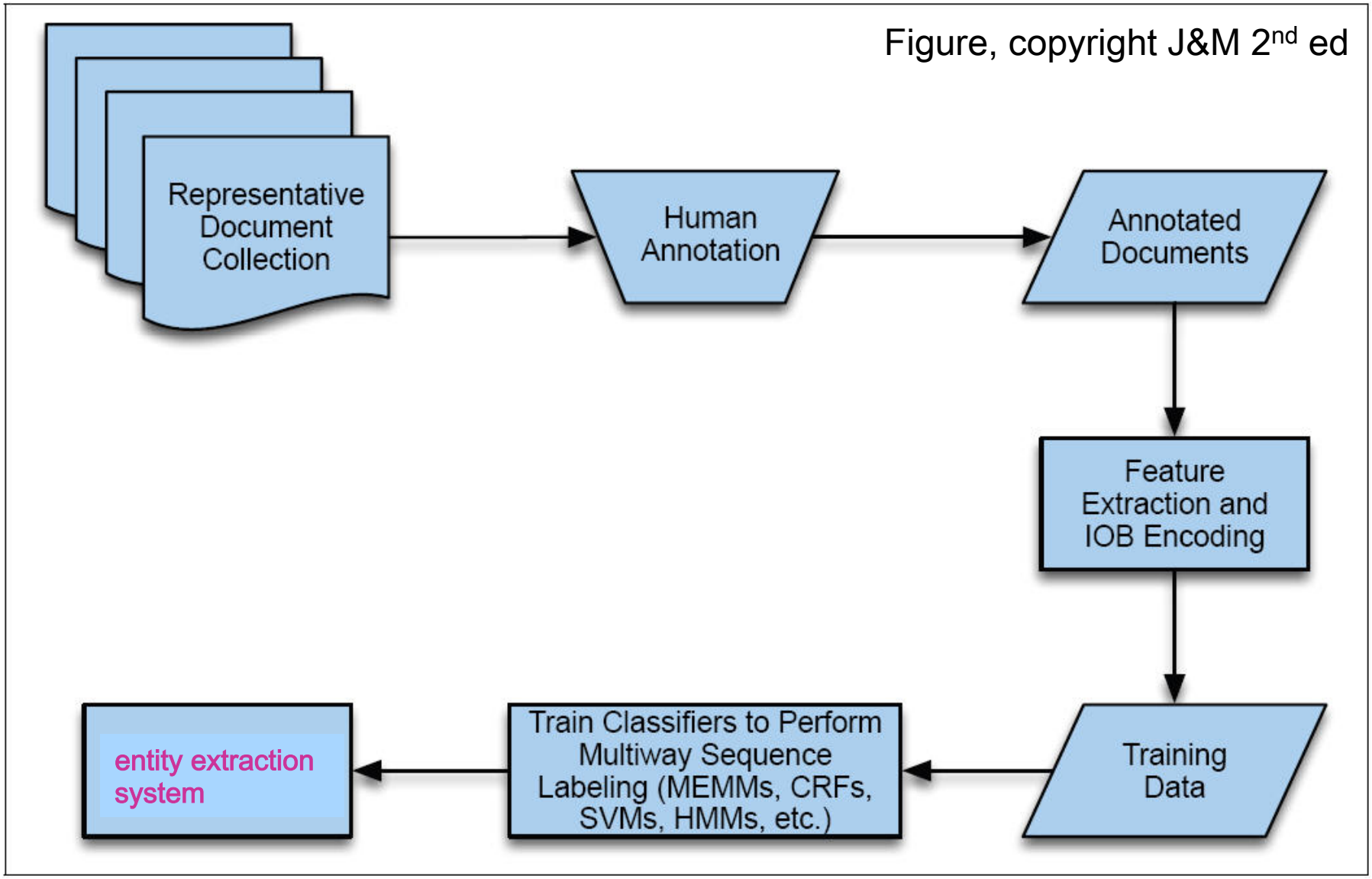
Window-based Classification

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



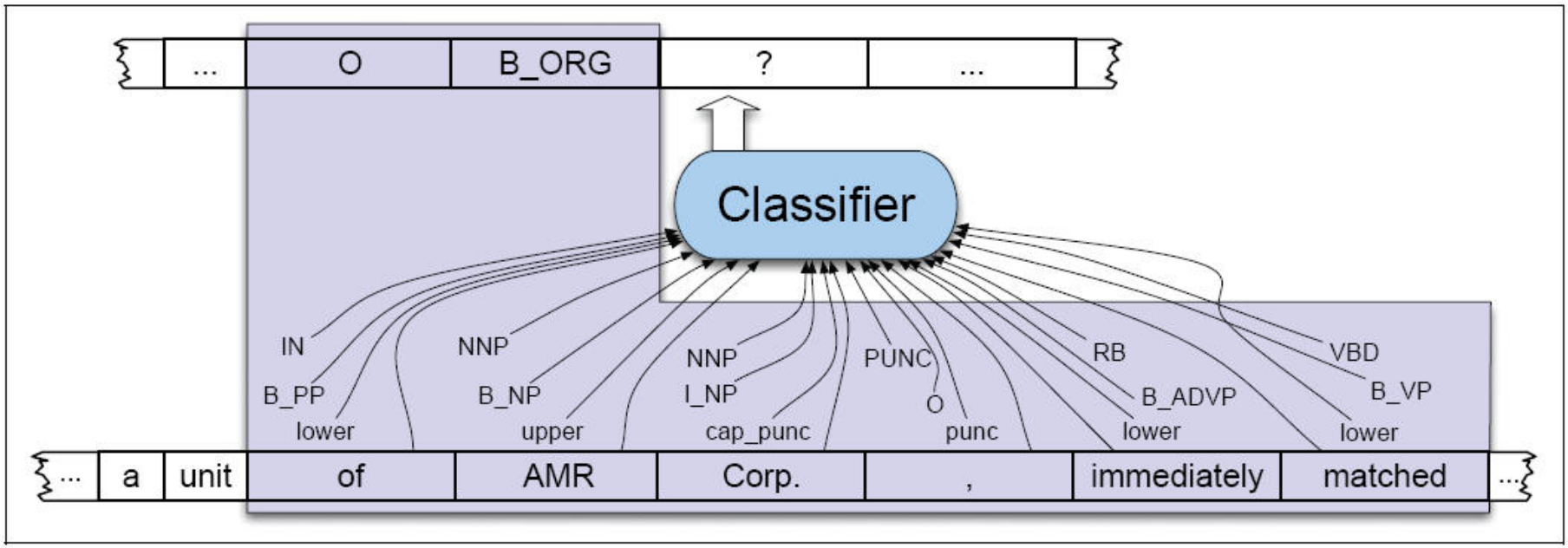
Figure, copyright J&M 2nd ed

End-to-end process



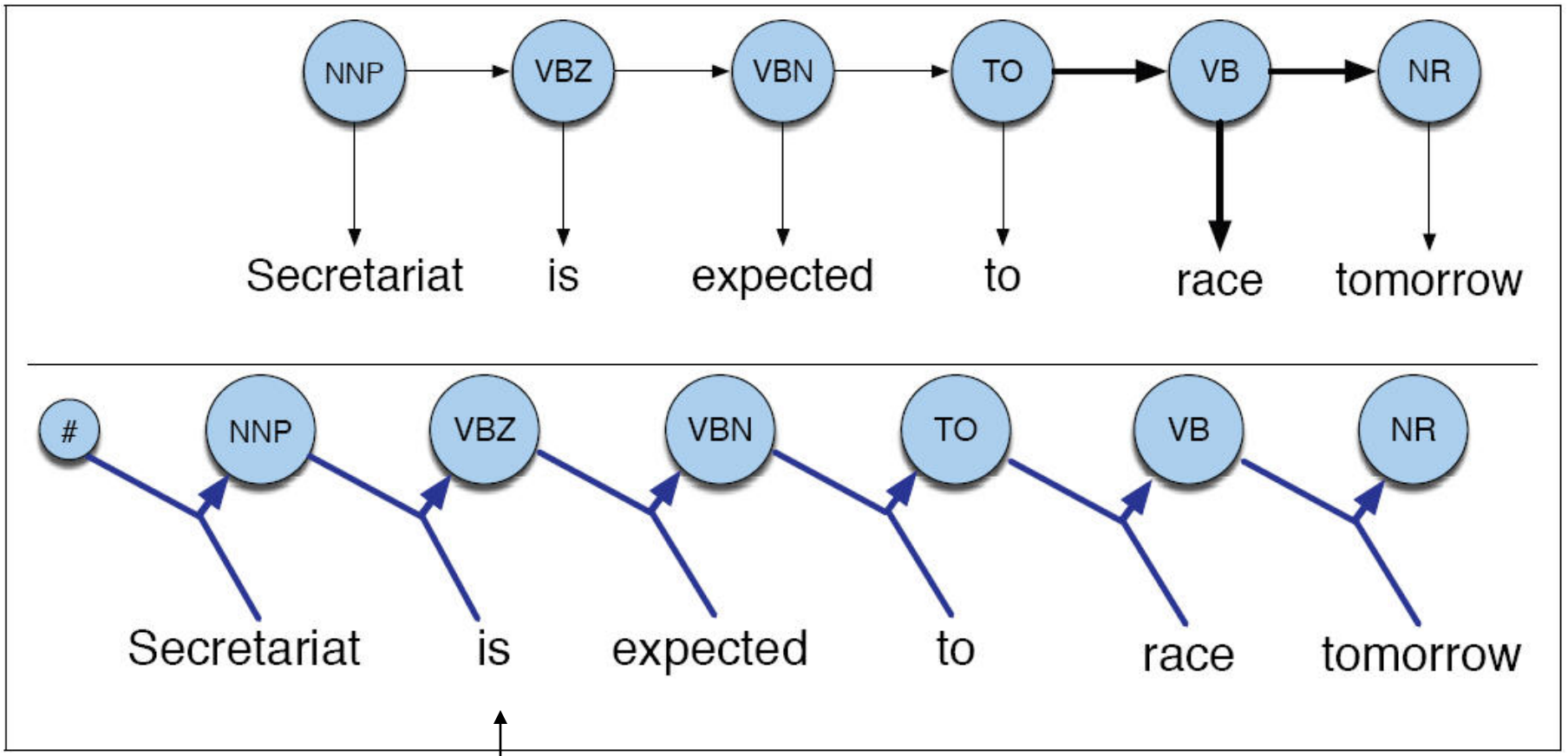
Feature extraction

- **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



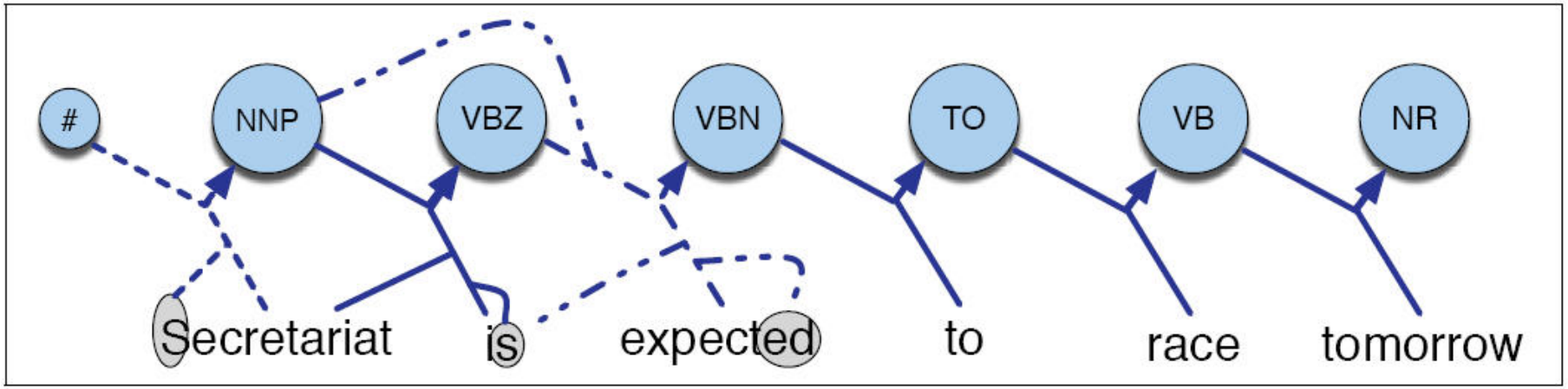
Figure, copyright J&M 2nd ed

MEMM equations

- **After spring break...**

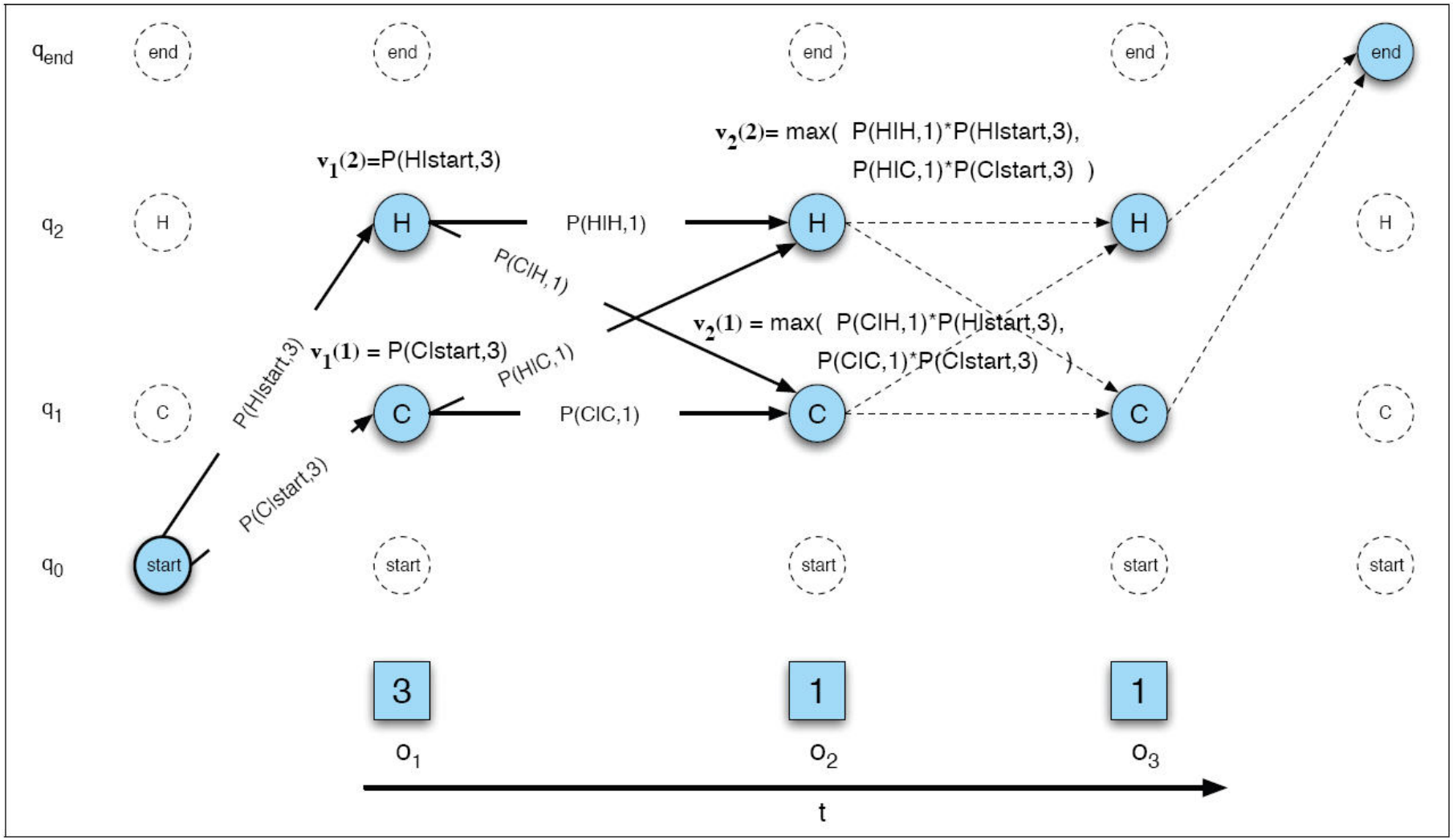
MEMM for p-o-s tagging

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



Figure, copyright J&M 2nd ed

Decoding/inference in MEMMs



Figure, copyright J&M 2nd ed

Information Extraction

- **Learning approaches**
  - Weakly supervised methods
  - Fully automatic methods for IE from structured text
  - Sequence-tagging methods
    - MEMM's
    - Opinion extraction

⇒ ILP for relation extraction

Relation extraction

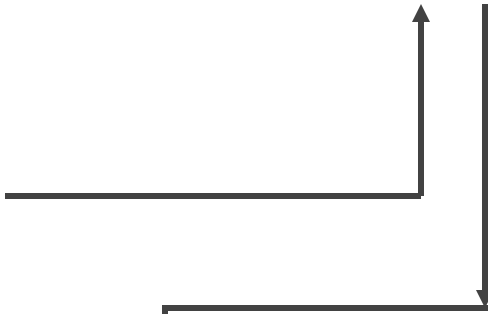
Relations		Examples	Types
Affiliations			
	Personal	*married to, mother of*	PER → PER
	Organizational	*spokesman for, president of*	PER → ORG
	Artifactual	*owns, invented, produces*	(PER | ORG) → ART
Geospatial			
	Proximity	*near, on outskirts*	LOC → LOC
	Directional	*southeast of*	LOC → LOC
Part-Of			
	Organizational	*a unit of, parent of*	ORG → ORG
	Political	*annexed, acquired*	GPE → GPE

Figure, copyright J&M 2nd ed

Fine-grained Opinions

"The Australian Press **launched a bitter attack** on Italy"

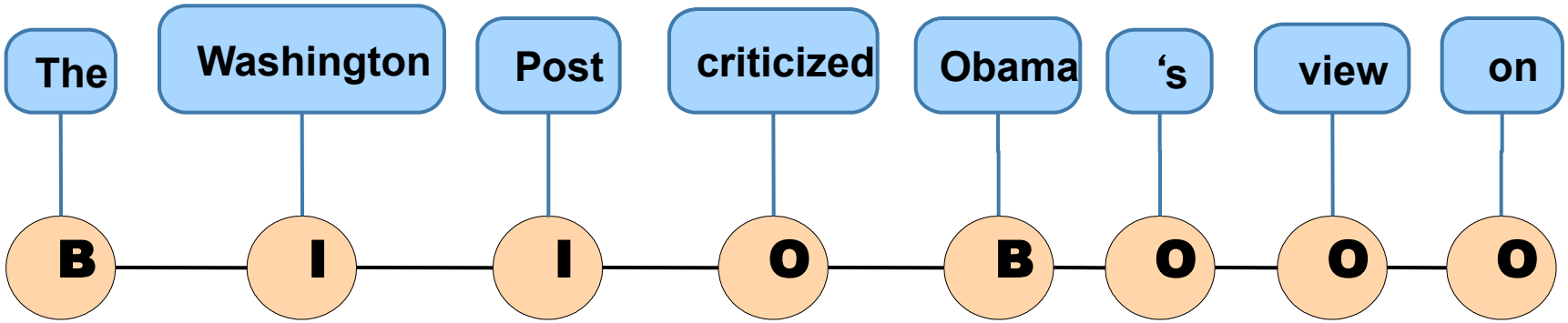
- **Five components**
  - Opinion trigger
  - Polarity
    - positive
    - negative
    - neutral
  - Strength/intensity
    - low..extreme
  - Source (opinion holder)
  - Target (topic)



Opinion Frame
Polarity: negative
Intensity: high
Source: "The Australian Press"
Target: "Italy"

Identifying Sources of Opinions

- **Via CRF's (extension of MEMM's)**



<The Washington Post> criticized **<Obama>**'s view on the oil crisis.

Features for Source Extraction

- Syntactically…

  - *mostly* noun phrases

- Semantically…

  - entities that can bear opinions

- Functionally…

  - linked to opinion expressions

Features for Source Extraction

- **Words [-4,+4]**
- **Capitalization**
- **Part-of-speech tags [-2,+2]**
- **Opinion phrase lexicon**
  - Derived from training data
  - Wiebe et al.'s [2002] 500+ word lexicon
- **Shallow semantic class information**
  - Sundance partial parser and named entity tagger
  - WordNet hypernym
- **Constituent type**
- **Grammatical role**
  - Collins' parser
- **Task-specific combinations**
  - E.g., Parent contains opinion word

Evaluation

- **MPQA data set** (www.cs.pitt.edu/mpqa)
  - ~550 documents
  - Manually annotated w.r.t. fine-grained opinion information
  - Provides gold standard
- **Automatically derive training/test examples**
- **10-fold cross-validation**
- **Evaluation measures**
  - Precision
  - Recall
  - F-measure

Results: Opinion Holders

>82% precision (accuracy)

~60% recall (coverage)

69.4 F-measure

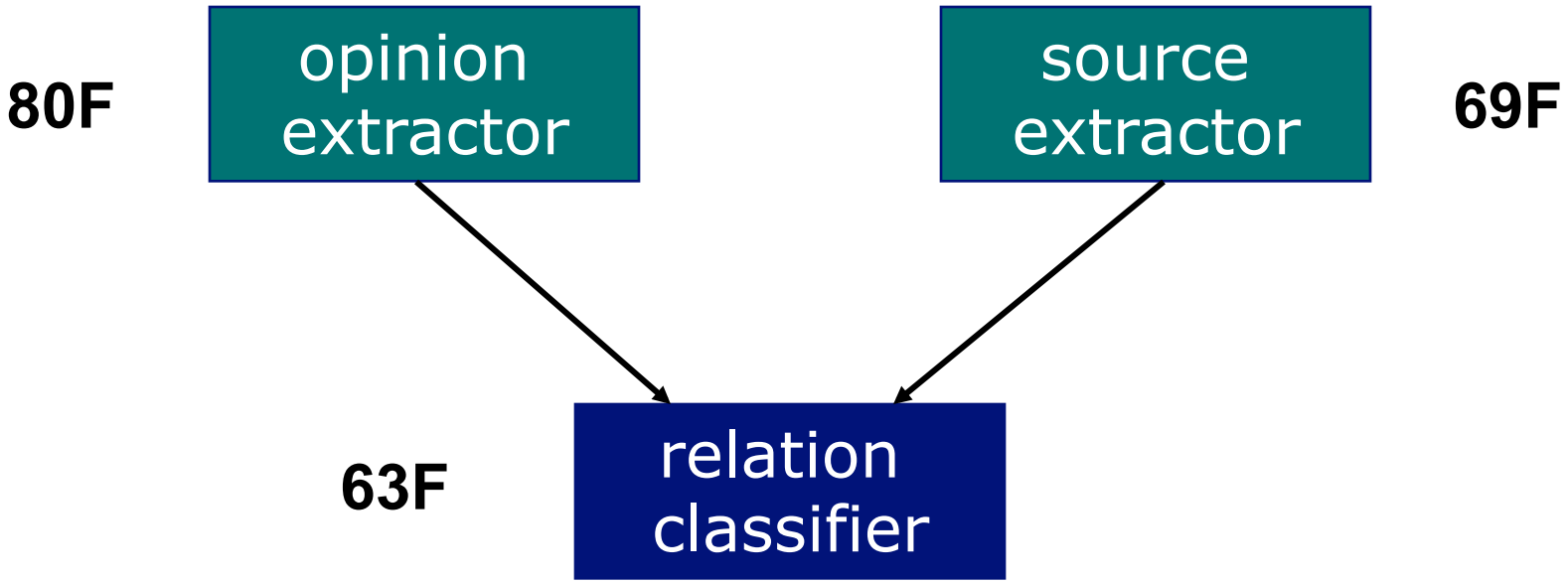
- **Better than a (very good!) pattern-learning IE approach** (Riloff)
- **Better than (very good!) semantic role labeling algorithms** (Roth)
- **But there's a lot of room for improvement…**

Errors

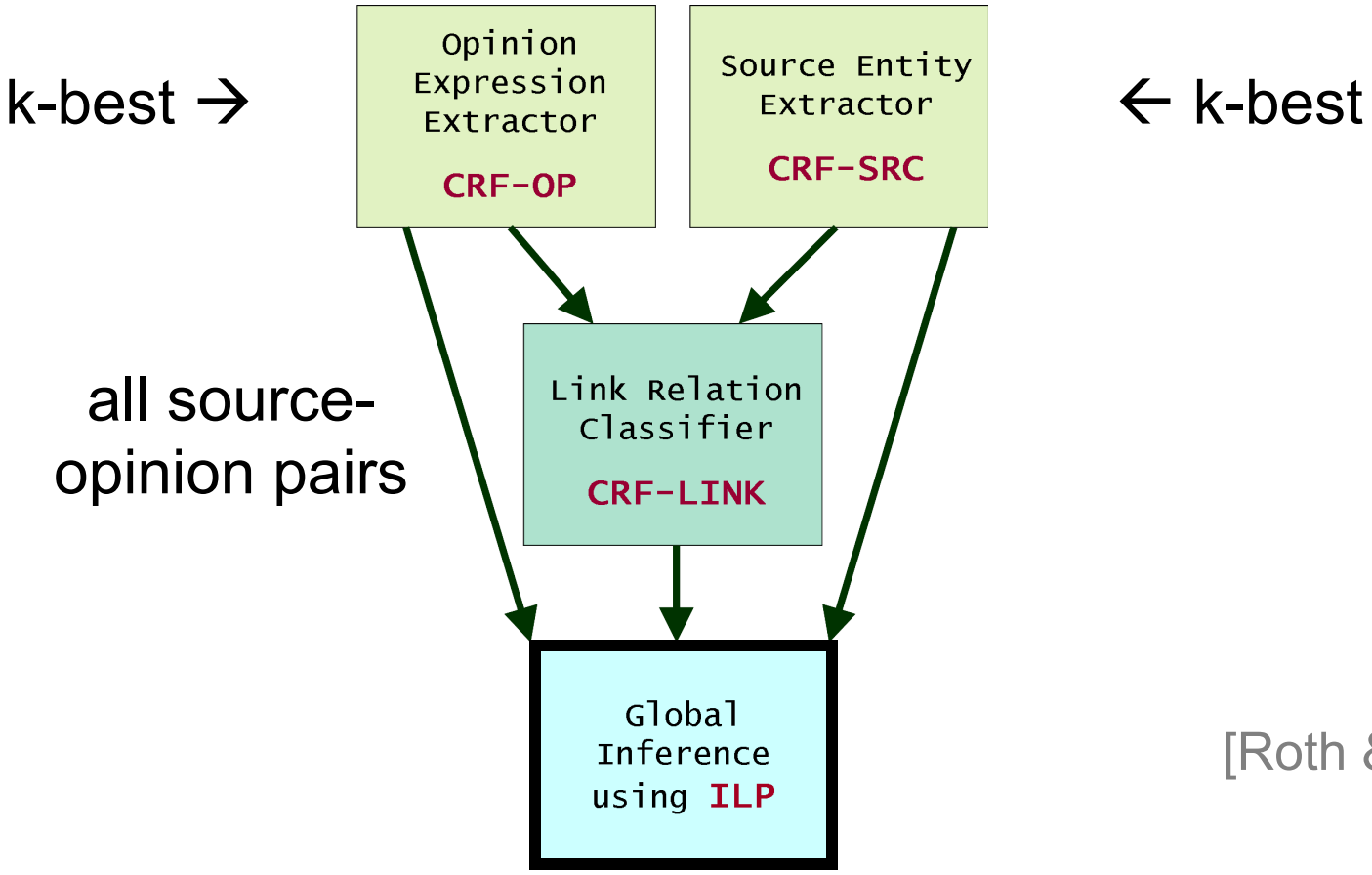
- **False positives**
  - Perhaps this is why **Fidel Castro has not spoken out** against what might go on in Guantanamo.
- **False negatives**
  - And for this reason, too, they have a moral duty to **speak out**, as **Swedish Foreign Minister Anna Lindh, among others**, did yesterday.
  - In particular, **Iran and Iraq are at loggerheads** with each other to this day.

Extracting and Linking to Opinions

- **To be useful, we need to link sources to their opinions**
  - <source> expresses <opinion>



Joint extraction of entities and relations



[Roth & Yih, 2004]

Constraints

- **Binary integer variables *O_i, S_j, L_i,j***
  - Weights for *O_i, S_j, L_i,j* are based on probabilities from individual classifiers
- **Constraints**

$$\forall i,\ O_i = \sum_j L_{i,j} \quad : \text{link coherency (only one link from each opinion)}$$

$$\forall j,\ S_j + A_j = \sum_i L_{i,j} \quad : \text{link coherency (upto two links from each source)}$$

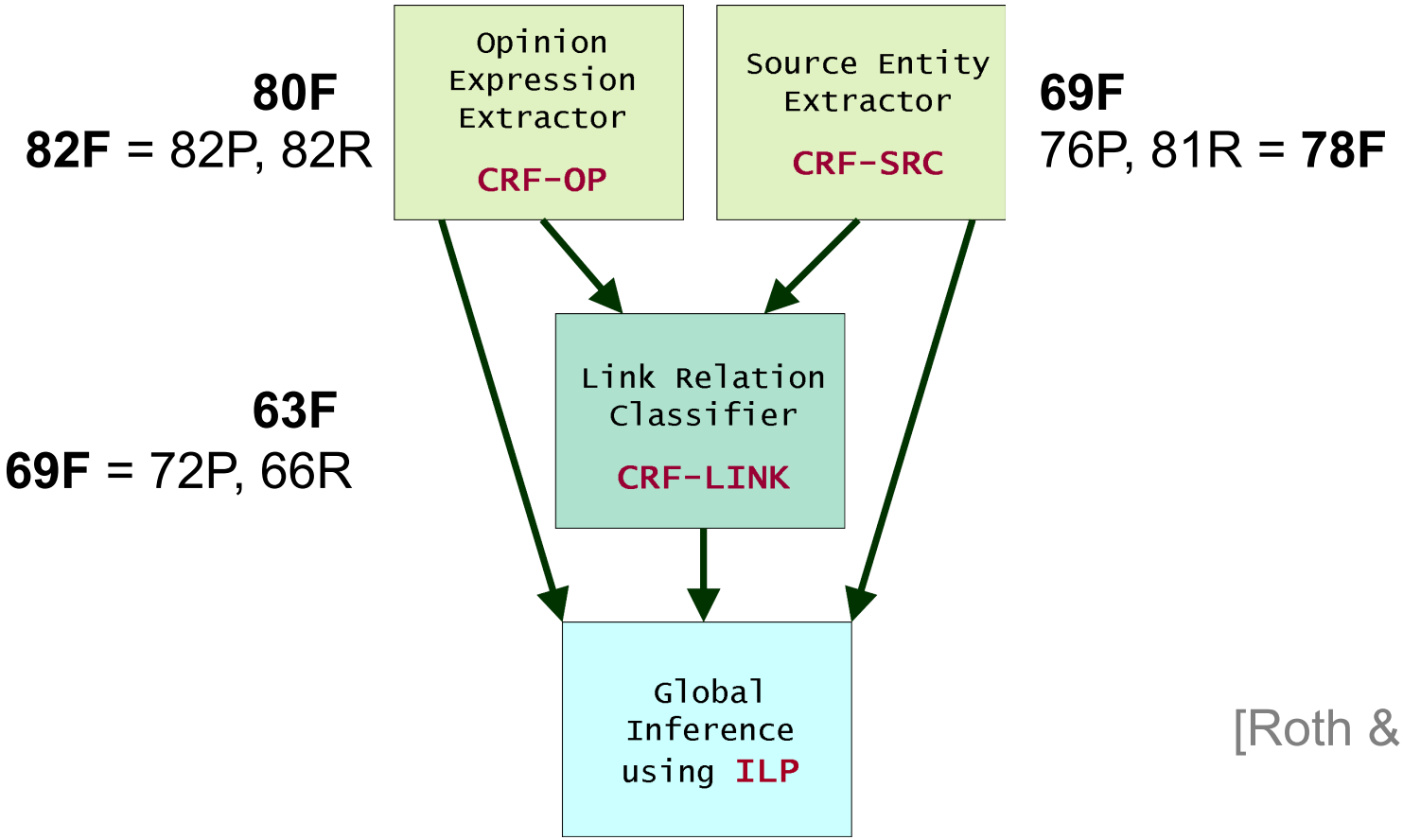
$$\forall j,\ A_j - S_j \le 0 \quad : \text{link coherency (preferably one link from each source)}$$

$$\forall i,j, i<j,\ X_i + X_j = 1,\ X \in \{S,O\}$$

- : entity coherency (for all pairs of entities with overlapping spans)

$$f = \sum_i (w_{O_i} O_i) + \sum_i (\overline{w}_{O_i} \overline{O}_i) + \sum_j (w_{S_j} S_j) + \sum_j (\overline{w}_{S_j} \overline{S}_j) + \sum_{i,j} (w_{L_{i,j}} L_{i,j}) + \sum_{i,j} (\overline{w}_{L_{i,j}} \overline{L}_{i,j})$$

Opinion Frame Extraction via CRFs and ILP



[Roth & Yih, 2004]