CS4740 Intro to NLP

- **Today: sequence tagging applications in NLP**
  - part-of-speech tagging
  - hidden Markov model (HMM)
  - named entity recognition (NER)
  - MEMMs

NE Identification

- **Identify all named locations, named persons, named organizations, dates, times, monetary amounts, and percentages.**

The delegation, which included the commander of the U.N. troops in Bosnia, Lt. Gen. Sir Michael Rose, went to the Serb stronghold of Pale, near Sarajevo, for talks with Bosnian Serb leader Radovan Karadzic.

Este ha sido el primer comentario publico del presidente Clinton respecto a la crisis de Oriente Medio desde que el secretario de Estado, Warren Christopher, decidiera regresar precipitadamente a Washington para impedir la ruptura del proceso de paz tras la violencia desatada en el sur de Libano.

1. Locations
2. Persons
3. Organizations

Figure 1.1 Examples. Examples of correct labels for English text and for Spanish text.

Guidelines need to be specified

- ***The Wall Street Journal*** **: artifact or organization?**
- ***White House*** **: organization or location?**
- **Is a street name a location?**
- **Should *yesterday* and *last Tuesday* be labeled as dates?**
- **Is *mid-morning* a time?**

Examples

1. **MATSUSHITA ELECTRIC INDUSTRIAL CO.** HAS REACHED AGREEMENT …
2. IF ALL GOES WELL, **MATSUSHITA** AND ROBERT BOSCH WILL …
3. **VICTOR CO. OF JAPAN** (**JVC**) AND SONY CORP. …
4. IN A FACTORY OF **BLAUPUNKT WERKE**, A **ROBERT BOSCH** SUBSIDIARY, …
5. **TOUCH PANEL SYSTEMS**, CAPITALIZED AT 50 MILLION YEN, IS OWNED …
6. **MATSUSHITA** EILL DECIDE ON THE PRODUCTION SCALE. …

Figure 2.1 English Examples. Finding names ranges from the easy to the challenging. Company names are in boldface. It is crucial for any name-finder to deal with the underlined text.

Training Data

- **Usually indicate NE's via SGML or XML**
  - Mark boundaries of expression
  - Label span with appropriate name class

Approaches to NE identification

- **Handcrafted finite state patterns**
  - <proper noun>$^{+}$ <corporate designator> → <corporation>
  - Can't easily capture typical naming conventions
    - "Boston Power & Light" (corporation, electric utility)
  - Time-consuming to define
  - Maintenance is a problem
    - E.g. moving to NYT from WSJ
  - Not generally portable to new languages

Identifinder [Bikel et al. 1997, 1999]

- **Hidden Markov model that learns to recognize and classify named entities.**
- **Outperforms other learning algorithms on standard data sets [MUC-6, MUC-7, MET-1]**
- **Competitive with approaches based on handcrafted rules on mixed case text**
- **Superior on text where case information isn't available**

Identifinder

- **Handles 7 classes of NE's**
  - entity
    - person
    - organization
    - location
  - time expression
    - date
    - time
  - numeric expression
    - money
    - percent

HMM's for NE identification

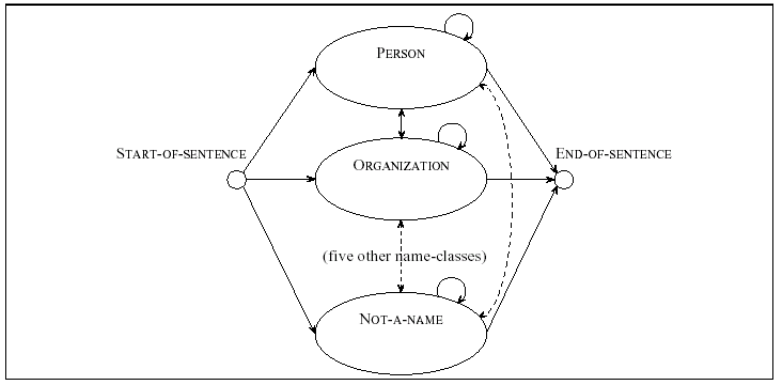
- **View NE identification as a sequence of word classification tasks**
- **Successful for other "word tagging" tasks, e.g. part-of-speech tagging**
- **Local cues to identify named entities**

- **Goal: Train an HMM to label every word with one of the NE name classes or with a *not-a-name* class.**

- **Alternative: MEMMs, CRFs …**

High-level view

A hidden Markov model represents the process of generating the sequence of words and labels



BBN's Identifinder (Bikel et al. 1999)

Using the HMM

- **Goal: find the most likely sequence of name classes, NC, given a sequence of words W**
  - W: *Banks filed bankruptcy papers*
  - Compare the probability of
    <person, not-a-name, not-a-name, not-a-name>
    <not-a-name, not-a-name, not-a-name, not-a-name>
    …
  - Viterbi algorithm is a dynamic programming algorithm that performs this computation efficiently.

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	IdentiFinder
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

CS4740 Intro to NLP

- **Today: sequence tagging applications in NLP**
  - part-of-speech tagging
  - hidden Markov model (HMM)
  - named entity recognition (NER)
  - MEMMs

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

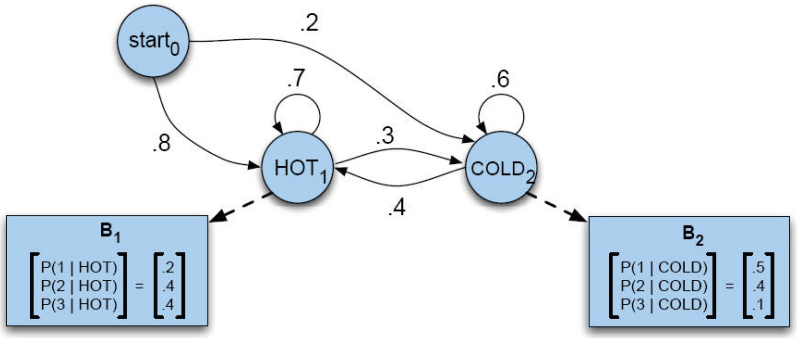
Figure, copyright J&M 2nd ed

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

Figure, copyright J&M 2nd ed

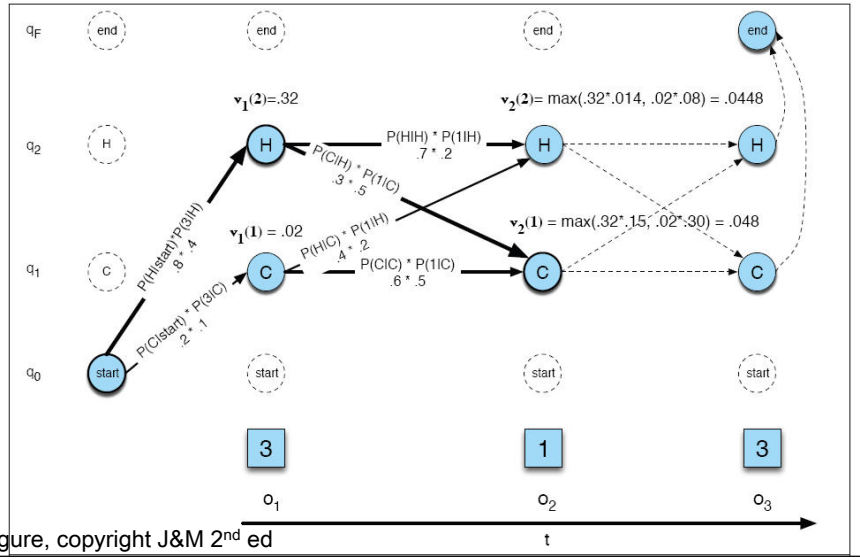
HMM for weather prediction



Figure, copyright J&M 2nd ed

HMM equations

Viterbi



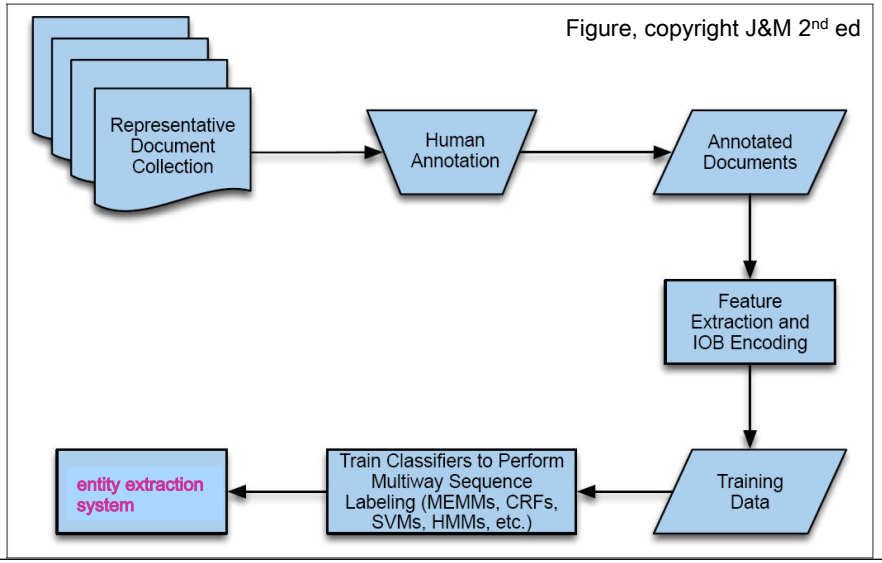
Figure, copyright J&M 2nd ed

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

Figure, copyright J&M 2nd ed

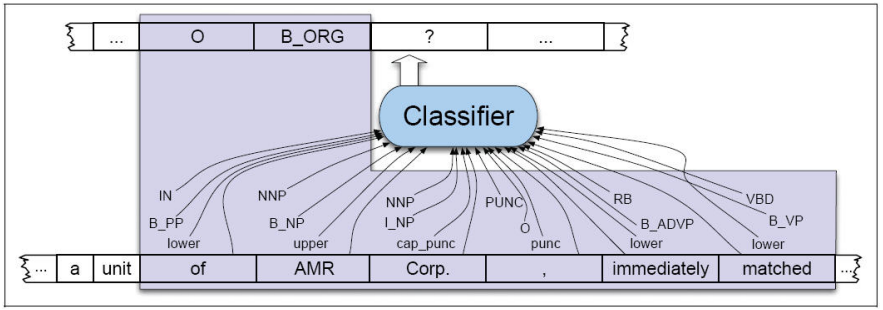
End-to-end process



Figure, copyright J&M 2nd ed

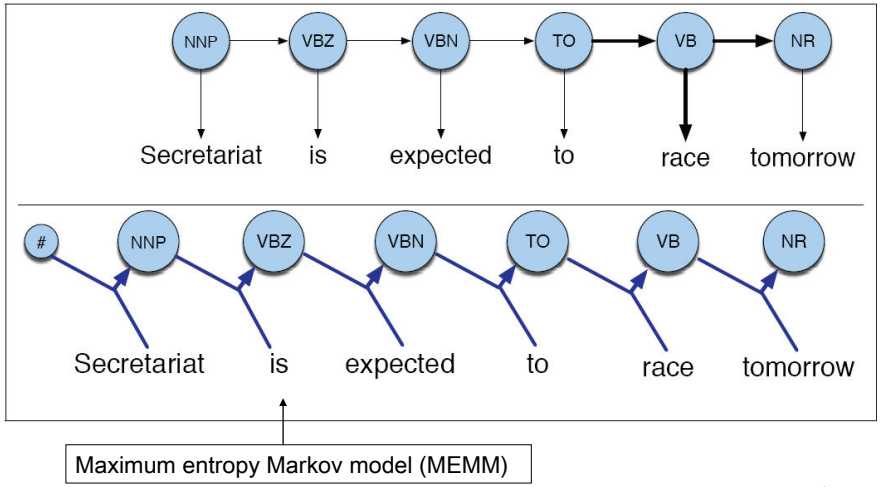
Feature extraction

- **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

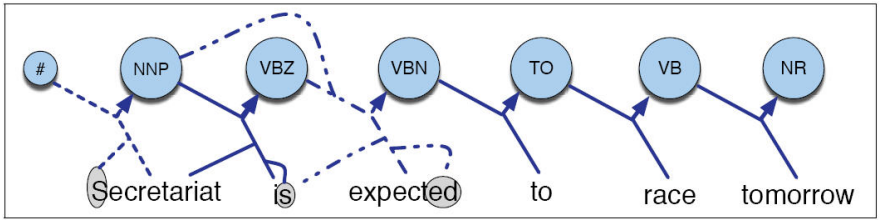


Figure, copyright J&M 2nd ed

MEMM equations

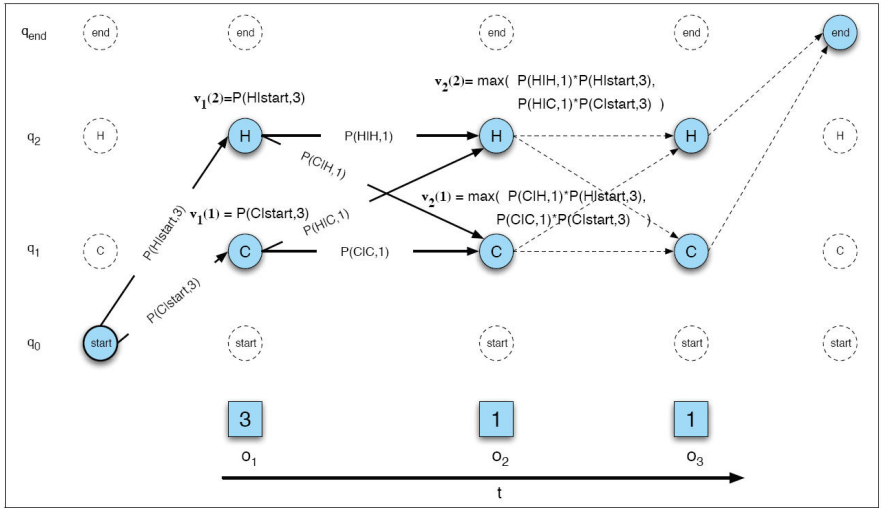
MEMM for p-o-s tagging

- **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

- **Next class**
  - Sentiment/opinion analysis