CS4740 Intro to NLP

- **Last classes: part-of-speech tagging**
  - part-of-speech tagging
  - hidden Markov model (HMM)
- **Today: another sequence tagging application in NLP**
  - named entity recognition (NER)
  - introduction to 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 NEs via SGML, XML, JSON**
  - 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

HMM's for NE identification

- View NE identification as a word-tagging task
  - 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 …

Identifinder [Bikel et al. 1997, 1999]

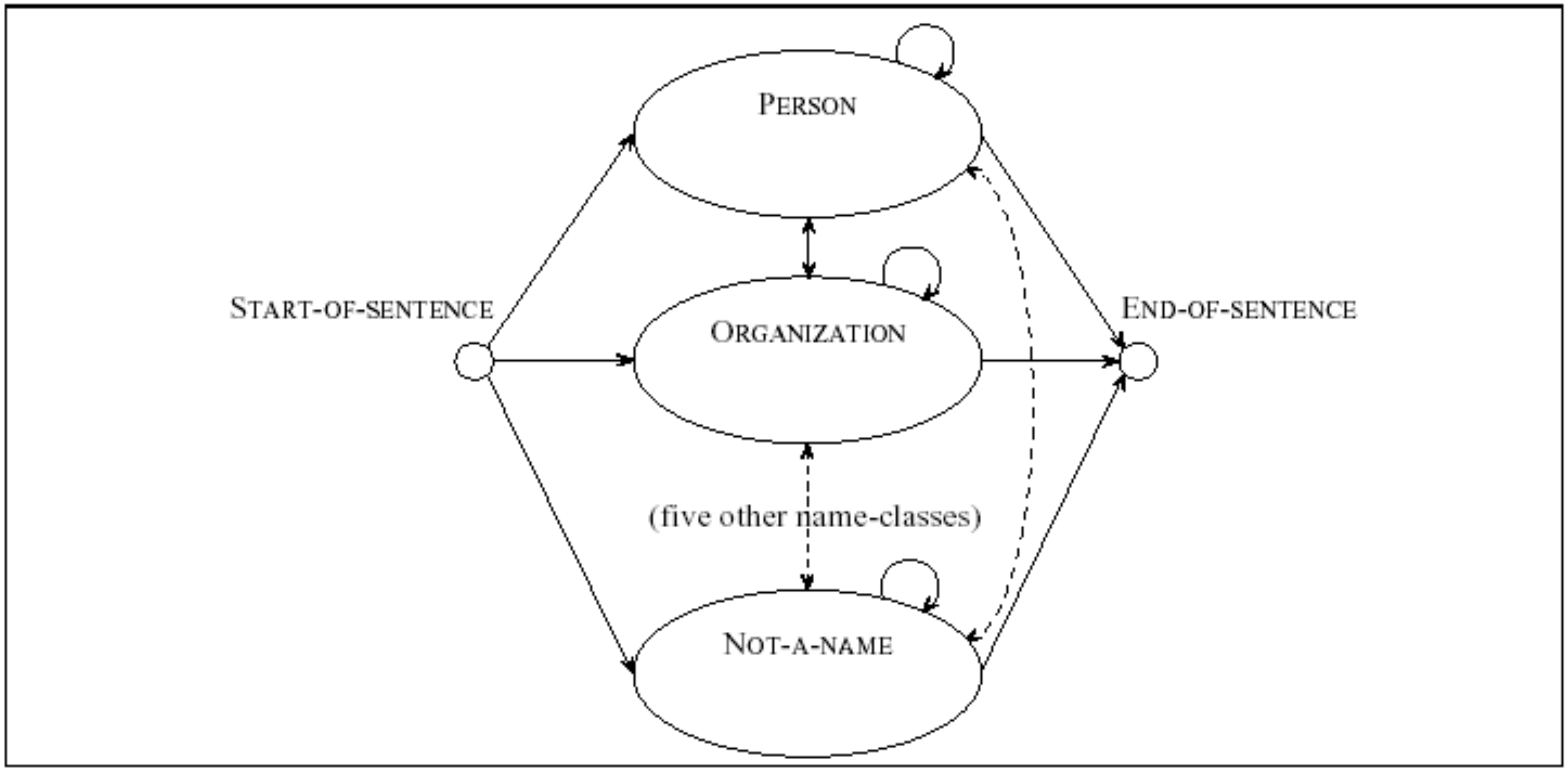
- First Hidden Markov model for recognizing and classifying named entities
- Outperformed other learning algorithms on standard data sets [MUC-6, MUC-7, MET-1]
- Competitive with approaches based on handcrafted rules on mixed case text
- Superior on text where case information isn't available

Identifinder

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

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)

States and transitions

- **States**
  - One for each name class
  - Special start and end states
- **Links have transition probabilities**
- **Each state also produces the words in each NE class (observables) based on**
  - the emission probability P(<word> | <state>)

Specifying the probabilities

- **Goal: Given a sequence of words W, find the sequence of name classes, NC, that maximizes P(NC|W)**
- **Restate using Bayes rule**
  - P(NC|W) = (P(NC) * P(W|NC)) / P(W)
- **Make independence assumptions**
  - Approximate each term

$$P(NC_0, NC_1, \ldots, NC_n) = \prod_{i=0}^{n} P(NC_i \mid NC_{i-1})$$

$$P(w_0, w_1, \ldots, w_n \mid NC_0, NC_1, \ldots, NC_n) = \prod_{i=0}^{n} P(w_i \mid NC_i)$$

Identifinder model

- used slightly different approximations

$$P(NC_0, NC_1, \ldots, NC_n) = \prod_{i=0}^{n} P(NC_i \mid NC_{i-1}, w_{i-1})$$

$$P(w_0, w_1, \ldots, w_n \mid NC_0, NC_1, \ldots, NC_n) = \prod_{i=0}^{n} P(w_i \mid NC_i, w_{i-1})$$

$$P(w_{first} \mid NC_i, NC_{i-1})$$

Using the HMM

- **Goal: find the most likely sequence of name classes, 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>

    …
  - As in HMMs for POS tagging, use the Viterbi algorithm.

Example

- **Computing the probability of a word-NC sequence:**
  - Mr. <name=person>Bill</name> talks.

P(not-a-name | start-of-sentence, +end+) *
    P("Mr." | not-a-name, start-of-sentence) *
P(person | not-a-name, "Mr.") *
    P("Bill" | person, not-a-name) *
P(not-a-name | person, "Bill") *
    P("talks" | not-a-name, person) *
P("." | "talks", not-a-name) *
    P(end-of-sentence | not-a-name, ".")

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