CS630 Representing and Accessing Digital Information

Named-Entity Recognition

Thorsten Joachims
Cornell University

Based on slides from Prof. Claire Cardie

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.

Data

- **Usually provided as SGML**
- **Marks boundaries of expression**
- **Labels span with appropriate name class**

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

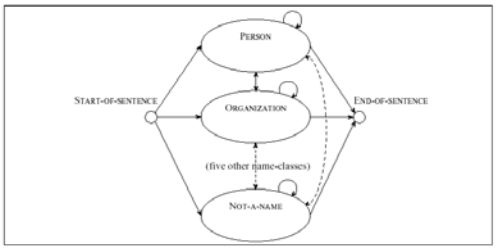
- **Handles 7 classes of NE's**
  - entity
    - person
    - organization
    - location
  - time expression
    - date
    - time
  - numeric expression
    - money
    - percent

Other Approaches to NE Identification

- **Previous techniques based on handcrafting 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**

High-Level View

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



States and Transitions

- **States (hidden)**
  - One for each name class
  - Two special start and end states
- **Links have transition probabilities**
- **Each (hidden) state also produces the words in each NE class (observables) based on**
  - prior probability of the state P(s)
  - the emission probability P(w|s)

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 using bigram model

    E.g. $P(NC_0, NC_1, ..., NC_n) = \prod_{i=0}^{n} P(NC_i \mid NC_{i-1})$
- **View each type of "name" to be its own language, with separate bigram probabilities for generating its words**

Components of the Identifinder Model

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

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

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

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

Feature Engineering

- **Use word-feature pairs instead of words**

$P(< w, f >_{first} | NC_i, NC_{i-1})$ $\quad$ $P(< w, f >_i | < w, f >_{i-1}, NC_i)$

- **Mutually exclusive features sort all words and punctuation into one of 14 categories**

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

Estimating the Probabilities

- **Estimate all of the necessary probabilities from the corpus**
  - E.g. $P(NC_i \mid NC_{i-1}, w_{i-1}) \approx$

    # times a word of name class NC_i follows word w_{i-1} of name class NC_{i-1} / total # of occurrences of word w_{i-1} with name class NC_{i-1}

- **Fairly complex back-off models and smoothing to handle sparse data.**

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 used for efficient decoding.

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, ".")

Results

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