CMPT-882: Statistical Learning of Natural Language

Lecture #6

Anoop Sarkar
anoop@cs.sfu.ca
http://www.sfu.ca/~anoop

- Nymble: a high-performance learning name-finder, Daniel M. Bikel and Scott Miller and Richard Schwartz and Ralph Weischedel. In *Proceedings of ANLP-97*, pages 194–201, 1997. `citeseer.nj.nec.com/bikel97nymble.html`

- An Algorithm that Learns What's in a Name, Daniel M. Bikel and Richard L. Schwartz and Ralph M. Weischedel. *Machine Learning*, volume 34, number 1-3, pages 211–231, 1999. `citeseer.nj.nec.com/bikel99algorithm.html`

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.

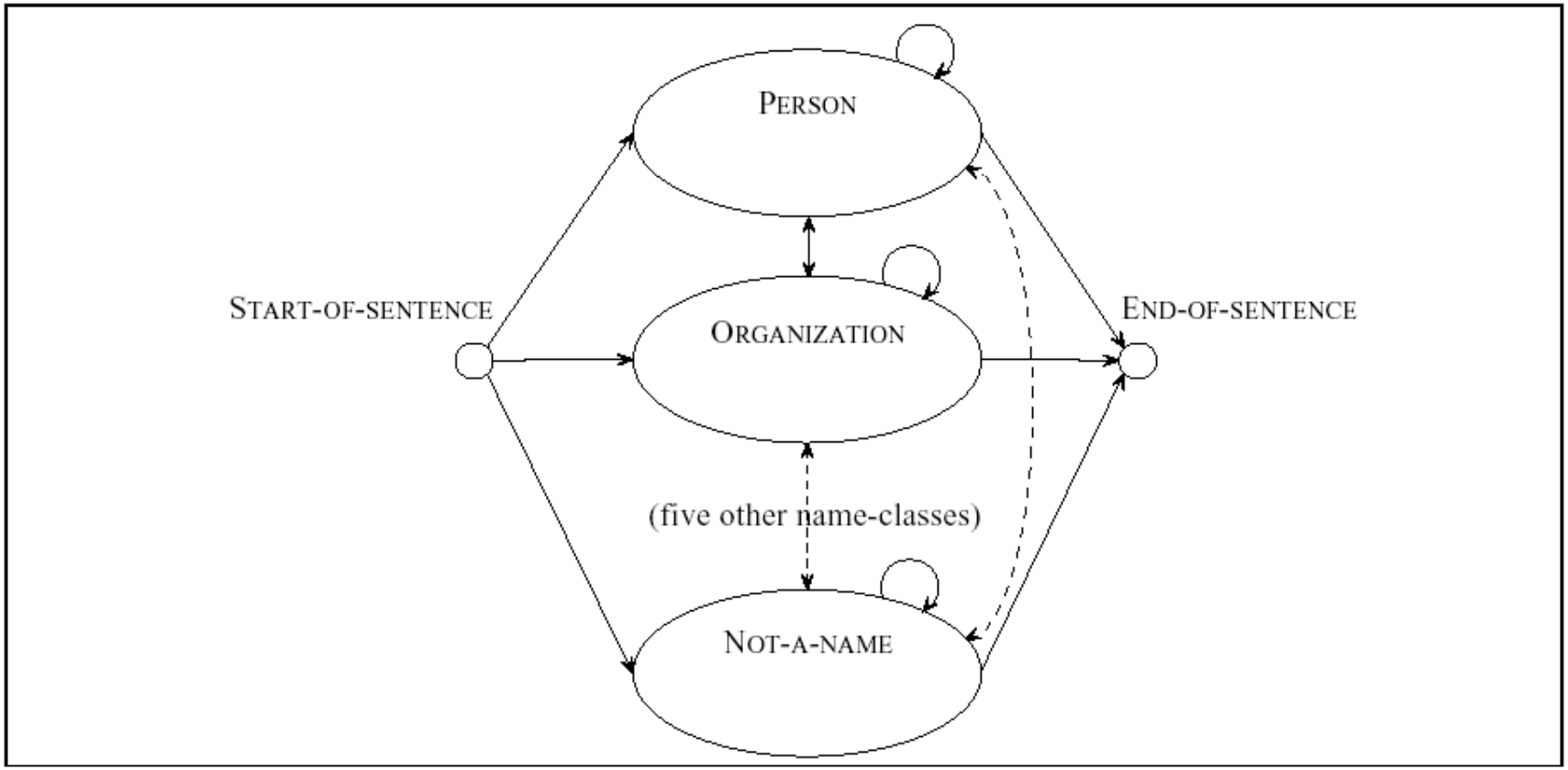


Figure 3.1 Pictorial representation of conceptual model. The subgraph of name-classes is complete, indicated here by the dashed arcs.

```
Mr. Jones eats.
```

According to rules of MUC and MET, the correct annotation for such a sentence is

```
Mr. <ENAMEX TYPE=PERSON>Jones</ENAMEX> eats.
```

That is, the token `Jones` is in the PERSON name-class, while the other tokens are in the NOT-A-NAME name-class. The model would assign the following likelihood to this word–name-class sequence (which we would hope to be the most likely, given sufficient training):

Pr(NOT-A-NAME | START-OF-SENTENCE, "+end+") *
Pr("Mr." | NOT-A-NAME, START-OF-SENTENCE) *
Pr(+end+ | "Mr.", NOT-A-NAME) *
Pr(PERSON | NOT-A-NAME, "Mr.") *
Pr("Jones" | PERSON, NOT-A-NAME) *
Pr(+end+ | "Jones", PERSON) *
Pr(NOT-A-NAME | PERSON, "Jones") *
Pr("eats" | NOT-A-NAME, PERSON) *
Pr("." | "eats", NOT-A-NAME) *
Pr(+end+ | ".", NOT-A-NAME) *
Pr(END-OF-SENTENCE | NOT-A-NAME, ".")

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

$$\Pr\left(NC \mid NC_{-1}, w_{-1}\right) = \frac{c(NC, NC_{-1}, w_{-1})}{c(NC_{-1}, w_{-1})}$$

$$\Pr\left(\langle w, f\rangle_{first} \mid NC, NC_{-1}\right) = \frac{c\left(\langle w, f\rangle_{first}, NC, NC_{-1}\right)}{c(NC, NC_{-1})}$$

$$\Pr\left(\langle w, f\rangle \mid \langle w, f\rangle_{-1}, NC\right) = \frac{c\left(\langle w, f\rangle, \langle w, f\rangle_{-1}, NC\right)}{c\left(\langle w, f\rangle_{-1}, NC\right)}$$

Table 3.2 Back-off strategy.

Name-class Bigrams	First-word Bigrams	Non–first-word Bigrams
$\Pr(NC \mid NC_{-1}, w_{-1})$	$\Pr(\langle w, f\rangle_{first} \mid NC, NC_{-1})$	$\Pr(\langle w, f\rangle \mid \langle w, f\rangle_{-1}, NC)$
$\vdots$	$\vdots$	$\vdots$
$\Pr(NC \mid NC_{-1})$	$\Pr(\langle w, f\rangle \mid \langle +begin+, \texttt{other}\rangle, NC)$	$\Pr(\langle w, f\rangle \mid NC)$
$\vdots$	$\vdots$	$\vdots$
$\Pr(NC)$	$\Pr(\langle w, f\rangle \mid NC)$	$\Pr(w \mid NC) \cdot \Pr(f \mid NC)$
$\vdots$	$\vdots$	$\vdots$
$\frac{1}{\text{number of name - classes}}$	$\Pr(w \mid NC) \cdot \Pr(f \mid NC)$	$\frac{1}{\lvert V\rvert} \cdot \frac{1}{\text{number of word features}}$
	$\vdots$	
	$\frac{1}{\lvert V\rvert} \cdot \frac{1}{\text{number of word features}}$	

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

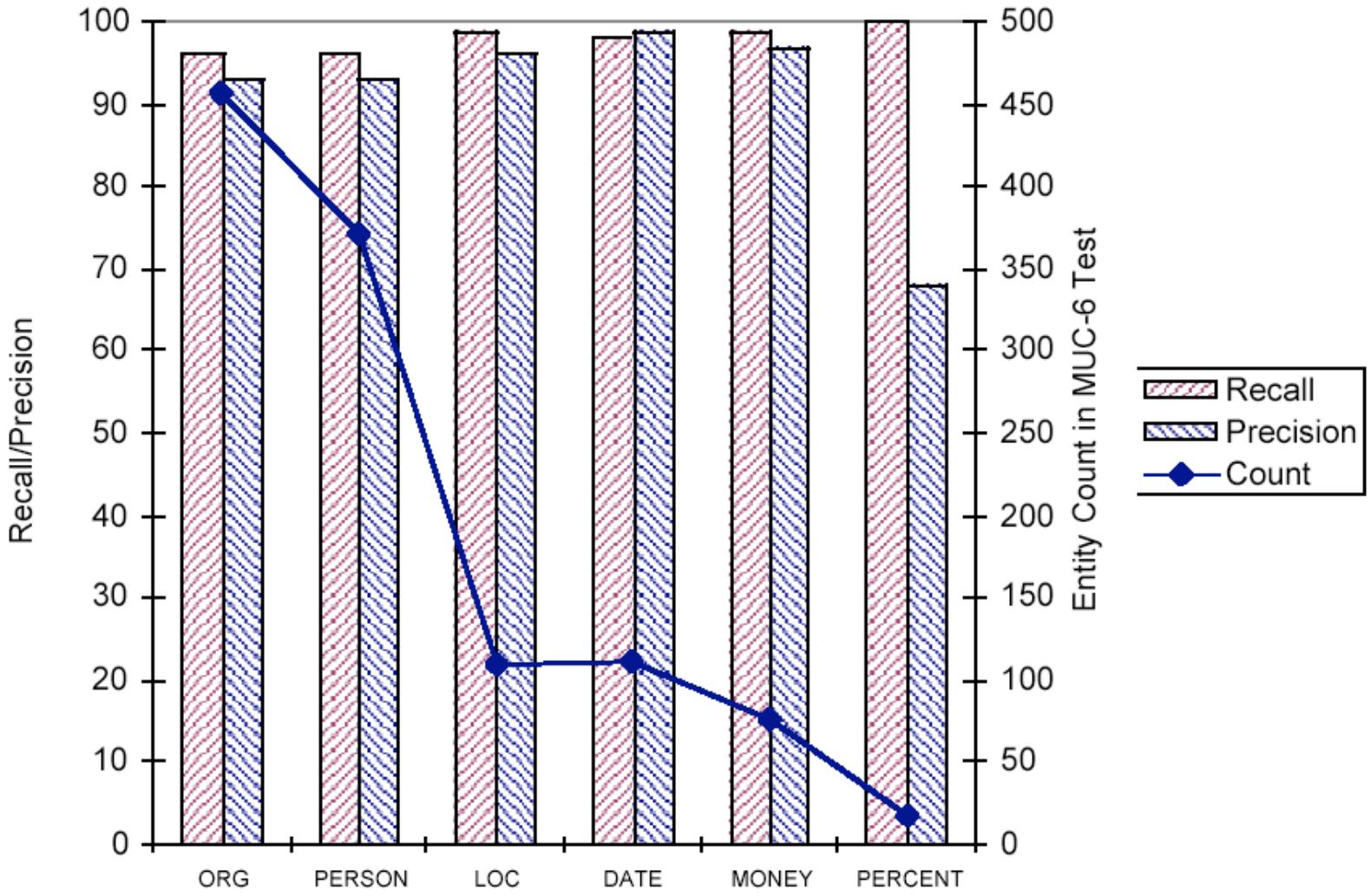
	Language	Best Rules	IdentiFinder
Mixed Case	**English (WSJ)**	**96.4**	**94.9**
Upper Case	**English (WSJ)**	**89**	**93.6**
Speech Form	**English (WSJ)**	**74**	**90.7**
Mixed Case	**Spanish**	**93**	**90**

Mixed Case: The British company, whose interests include the Cunard cruise lines and the Ritz hotel of London, said Simon Keswick, the head of Hong Kong Land Holdings Ltd. will take over as chairman May 26. The current acting chairman, Alan, will become a deputy chairman of the group.

Upper Case: THE BRITISH COMPANY, WHOSE INTERESTS INCLUDE THE CUNARD CRUISE LINES AND THE RITZ HOTEL OF LONDON, SAID SIMON KESWICK, THE HEAD OF HONGKONG LAND HOLDINGS LTD. WILL TAKE OVER AS CHAIRMAN MAY 26. THE CURRENT ACTING CHAIRMAN, ALAN CLEMENTS, WILL BECOME A DEPUTY CHAIRMAN OF THE GROUP.

SNOR: THE BRITISH COMPANY WHOSE INTERESTS INCLUDE THE CUNARD CRUISE LINES AND THE RITZ HOTEL OF LONDON SAID SIMON KESWICK THE HEAD OF HONGKONG LAND HOLDINGS LIMITED WILL TAKE OVER AS CHAIRMAN MAY TWENTY SIX THE CURRENT ACTING CHAIRMAN ALAN CLEMENTS WILL BECOME A DEPUTY CHAIRMAN OF THE GROUP

Figure 5.2 Three Modalities. The task becomes increasingly difficult as one moves from mixed case to upper case, to SNOR format.


Recall/Precision
Entity Count in MUC-6 Test
100
90
80
70
60
50
40
30
20
10
0
500
450
400
350
300
250
200
150
100
50
0
ORG
PERSON
LOC
DATE
MONEY
PERCENT
Recall
Precision
Count


100
95
90
85
80
75
F-measure
10000
100000
1000000
Training Data Size (word count)
English
Spanish