NYU

Evaluation

CSCI-GA.2590 – Lecture 6A

Ralph Grishman

Measuring Performance

- NLP systems will be based on models with simplifying assumptions and limited training, so their performance will never be perfect
  - to be able to improve the systems we build, we need to be able to measure the performance of individual components and the entire system

Accuracy

- for part-of-speech tagging, accuracy is a simple and reasonablemetric

- $\text{accuracy} = \dfrac{\text{tokens with correct tag}}{\text{total tokens}}$

Accuracy can be Misleading

- For tasks where one tag predominates, accuracy can overstate performance
- Consider name tagging for texts where 10% of the tokens are names
- A ‘baseline’ name tagger which tags every token as ‘other’ (not a name) would be rated as 90% accurate though it finds no names

Precision and Recall

Instead of counting the tags themselves, we count the names defined by these tags:

key = number of names in key

response = number of names in system response

correct = number of names in response which exactly match (in type and extent) a name in the key

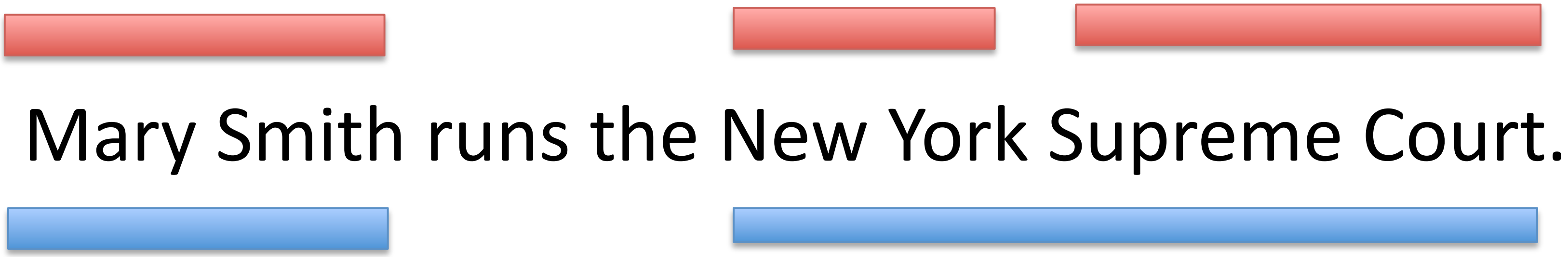
then

precision = correct / response

recall = correct / key

Precision & Recall (Example)

NE system response = 3



Mary Smith runs the New York Supreme Court.

NE key = 2

NE correct = 1

recall = 50%

precision = 33%

F-measure

We sometimes want a single measure to compare systems

The usual choice is F-measure, the harmonic mean of recall and precision

$$\frac{1}{F} = \frac{1}{2}\left(\frac{1}{precision} + \frac{1}{recall}\right)$$

$$F = 2 \times \frac{precision \times recall}{precision + recall}$$

Honest Test Data

For honest evaluations, test data should remain 'blind'

- avoid training to the test
- for a corpus-trained system, set aside separate test data

Cross-Validation