Precision and Recall
(based on Jurafsky and Martin)

Miriam Butt
January 2013
Evaluation

How can the performance of a system be evaluated?

Standard Methodology from Information Retrieval:

- Precision
- Recall
- F-measure (combination of Precision/Recall)
Evaluation

Get a reference corpus and use it as a “Gold Standard”

This Gold Standard is usually annotated manually for whatever application is being targeted (POS-tagging, parsing, semantic annotation).

See how well the system performs with respect to the Gold Standard.
Recall

Measure of how much relevant information the system has extracted (coverage of system).

Recall = \frac{\text{# of correct answers given by system}}{\text{total # of possible correct answers in text}}
Precision

Measure of how much of the information the system returned is correct (accuracy).

Precision = \frac{\text{# of correct answers given by system}}{\text{# of answers given by system}}
F-measure

Precision and Recall stand in opposition to one another. As precision goes up, recall usually goes down (and vice versa).

The F-measure combines the two values.

\[
F\text{-measure} = \frac{(\beta^2+1)PR}{\beta^2 P+R}
\]

- When \( \beta = 1 \), precision and recall are weighted equally.
- When \( \beta \) is > 1, precision is favored.
- When \( \beta \) is < 1, recall is favored.