Text Categorization: Evaluation

Part 1

ChengXiang "Cheng" Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign

Overview

- What is text categorization?
- Why text categorization?
- How to do text categorization?
 - Generative probabilistic models
 - Discriminative approaches
- How to evaluate categorization results?

General Evaluation Methodology

- Have humans to create a test collection where every document is tagged with the desired categories ("ground truth")
- Generate categorization results using a system on the test collection
- Compare the system categorization decisions with the human-made categorization decisions and quantify their similarity (or equivalently difference)
 - The higher the similarity is, the better the results are
 - Similarity can be measured from different perspectives to understand the quality of results in detail (e.g., which category performs better?)
 - In general, different categorization mistakes may have a different cost that inevitably depends on specific applications, but it is okay not to consider such a cost variation for relative comparison of methods

Classification Accuracy (Percentage of Correct Decisions)

 $\mathbf{d_N}$...

Classification Accuracy =

Total number of correct decisions

Total number of decisions made

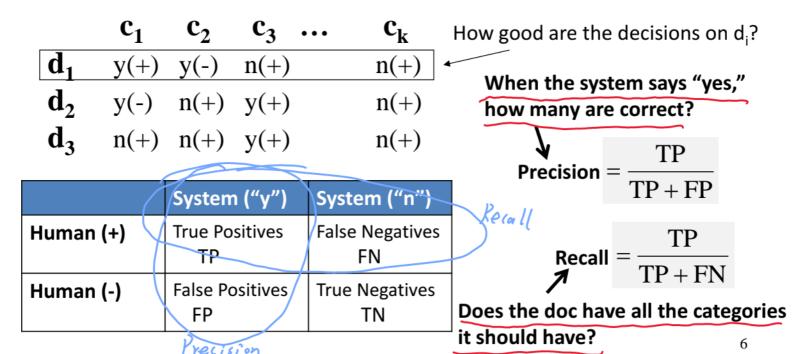
$$\frac{count(y(+)) + count(n(-))}{kN}$$

Problems with Classification Accuracy

- Some decision errors are more serious than others
 - It may be more important to get the decisions right on some documents than others
 - It may be more important to get the decisions right on some categories than others
 - E.g., spam filtering: missing a legitimate email costs more than letting a spam go
- Problem with imbalanced test set
 - Skewed test set: 98% in category 1; 2% in category 2
 - Strong baseline: put all instances in category 1 → 98% accuracy!

5

Per-Document Evaluation



Per-Category Evaluation

	System ("y")	System ("n")
Human (+)	True Positives TP	False Negatives FN
Human (-)	False Positives FP	True Negatives TN

How good are the decisions on c_i ?

When the system says "yes," how many are correct?

Precision
$$= \frac{TP}{TP + FP}$$

$$\mathbf{Recall} = \frac{\mathbf{TP}}{\mathbf{TP} + \mathbf{FN}}$$

Has the category been assigned to all the docs of this category?

Combine Precision and Recall: F-Measure

$$F_{\beta} = \frac{1}{\frac{\beta^{2}}{\beta^{2}+1} \frac{1}{R} + \frac{1}{\beta^{2}+1} \frac{1}{P}} = \frac{(\beta^{2}+1)P * R}{\beta^{2}P + R}$$

$$F_1 = \frac{2PR}{P + R}$$

Why not 0.5*P+0.5*R?

P: precision **R**: recall

β: parameter (often set to 1)

What is R if the system says "y" for all category-doc pairs?