Chapter 8:
Evaluation Measures

Overview

- The concept of relevance in information retrieval
- Evaluation measures:
  - information retrieval
  - subtasks of retrieval: text categorization, information extraction, summarization, ...
- Test collections:
  - information retrieval
  - subtasks of retrieval: text categorization, information extraction, summarization, ...

Relevance

- Plays a crucial role in the evaluation of the information retrieved, but difficult to express in exact numbers
- Relevance in text retrieval has different facets:
  - **topical relevance**: the subjects of a text
  - **motivational relevance** and **interpretational relevance***:*
    - the purpose of the search
    - the intended use of the information
    - the background of the user
    - informativeness of the information

Relevance

- Problems in text retrieval:
  - the natural language understanding of document texts and the user's preferences
  - the dynamic nature of the information need (even within a single retrieval session !)
  - -> impossible: in all circumstances to identify precisely and completely the subset of documents relevant to a given user in the context of a specific need
- In text retrieval: **weaker notions of relevance**:
  - relevance is the property of a document's being potentially helpful to a user in the resolution of a need
  - topical relevance is a necessary (first filter), but not sufficient condition for relevance

Evaluation measures: information retrieval

- Common criteria:
  - execution efficiency: time of computations of search and maintenance operations
  - storage efficiency: often measured by the space overhead
  - functionality for the user
  - **retrieval effectiveness** (possibly including the effectiveness of subtasks):
    - e.g., precision, recall, …
    - credibility of the information (e.g., on the World Wide Web)
  - -> **allows comparing technologies and systems**

Evaluation measures: information retrieval

- For a given query *q* a number of documents are retrieved: we can measure the recall and precision of the results

$$\text{recall} = \frac{|Arel|}{|Rel|} \qquad \text{precision} = \frac{|Arel|}{|A|}$$

where

Rel = set of relevant documents

A = set of documents that forms the answer list

Arel = set of relevant documents in *A*

Evaluation measures: information retrieval

- Recall and precision:
  - are measured for a given query *q* and possibly averaged over a set of queries *Q*
  - ideally close to 1
- Recall versus precision graph
  - computation of precision is based on 11 standard recall levels: 0%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100%
  - often interpolation needed to compute intermediate values between given values (and to compute 0% level)
  - recall and precision are often inversely related
  - for several queries: average of the precision figures at each recall level

Recall versus precision graph: example

$Rel_q = \{d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{71}, d_{89}, d_{123}\}$

System retrieves and ranks following documents (* = relevant):

1. d_{123}*
2. d_{84}
3. d_{56}*
4. d_6
5. d_8
6. d_9*
7. d_{511}
8. d_{129}
9. d_{187}
10. d_{25}*
11. d_{38}
12. d_{48}
13. d_{250}
14. d_{113}
15. d_3*

d_{123} gives 100% precision at 10% recall

d_{56} gives 66.7% precision at 20% recall …

Recall versus precision graph: example

Suppose $Rel_q = \{d_3, d_{56}, d_{129}\}$

d_{56} gives 33.3% precision at 33.3% recall

d_{129} gives 25% precision at 66.7% recall

d_3 gives 20% precision at 100% recall

The precision figures at the other standard recall levels are interpolated as follows:

$$P(r_j) = \max_{r_j \leq r \leq r_{j+1}} P(r)$$

where r_j references the j-th standard recall level

here: at 0%, 10%, 20%, 30% recall: precision is 33.3%

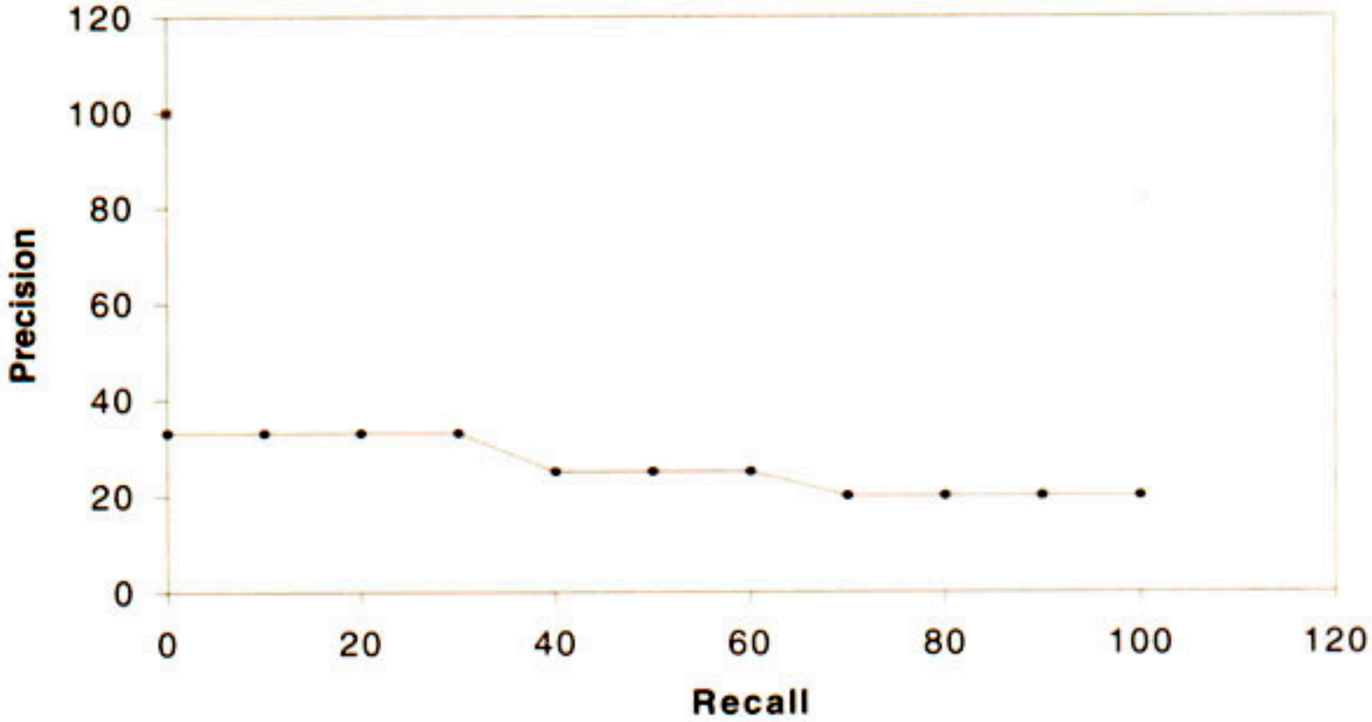


Figure 3.3 Interpolated precision at 11 standard recall levels relative to $R_q = \{d_3, d_{56}, d_{129}\}$.

source: Baeza-Yates & Ribeiro-Neto

Evaluation measures: information retrieval

- Recall and precision at a given document **cutoff** value λ: e.g., 10, 50 documents

- ***R*-precision**: precision (or recall) at the Rth position in the ranking where $R = |Rel|$, i.e., the number of relevant documents for the query

- **Breakeven point**: point in the recall versus precision graph where recall equals precision

Evaluation measures: information retrieval

- **F-measure**: combines recall and precision

$$F = \frac{(\beta^2 + 1)\ \text{precision x recall}}{\beta^2\ \text{precision} + \text{recall}}$$

where

β = a factor that indicates the relative importance of recall and precision

- ideally close to 1
- when $\beta = 1$: also called harmonic mean= F_1

Evaluation measures: information retrieval

- When retrieval from large databases (e.g., World Wide Web): concern for a high precision, especially when considering the top ranked items in the answer list
- The user is interested in receiving to-the-point (perhaps smallest retrieval element) that answers the information need

Evaluation measures: information retrieval

- **Non-interpolated average precision** (***AP***):

$$AP = \frac{1}{|Rel|} \sum_{r=1}^{|Rel|} P_r$$

$$P_r = \frac{|Arel_r|}{|A_r|}$$

where $Arel_r$ = the set of relevant documents of the answer list up to the position of the r^{th} relevant document; A_r is the set of retrieved documents up to position of the r^{th} relevant document; if the r^{th} relevant document does not occur in the answer list, $P_r = 0$

- When averaged over a set of queries: **mean average precision** (***MAP***)

Evaluation measures: information retrieval

- **Preferences:**
  - use preferences from binary relevance judgments by comparing preferences generated by a system with the preferences generated by an expert
- **Binary preference (*bpref*):**
  - a given query has *R* relevant documents
  - we consider up to *R* non-relevant documents in the answer list
  - => *R* x *R* preference judgments (for the expert all relevant documents are preferred above all non relevant documents)

Evaluation measures: information retrieval

- **bpref** = $\frac{P}{P+Q}$
  - where P = number of preference agreements and Q = number of preference non-agreements ($P + Q = R \times R$)
- Practically **bpref** is computed as: $\frac{1}{R}\sum_{r=1}^{R}(1 - \frac{N_r}{R})$
  - where N_r = the number of non-relevant documents (from the set of R non-relevant documents that are considered) that are ranked higher than the relevant document at rank r

Evaluation measures: information retrieval

- **Discounted cumulative gain** (DCG):
  - Computed as the total gain accumulated at a particular rank *r*

$$DCG_r = rel_1 + \sum_{i=2}^{r} \frac{rel_i}{\log_2 i}$$

  - where rel_i = graded relevance level or binary relevance of the document retrieved at rank *i*
- **Graded relevance level**:
  - e.g., six point scale from “bad” to “perfect” ($0 <= rel_i <= 5$)
  - often used in **Web search evaluations**

Evaluation measures: information retrieval

- **Mean Reciprocal Answer Rank (*MRAR*):** used in question answering retrieval for evaluating a set of queries

$$MRAR = \frac{1}{n}\left(\sum_{i=1}^{n} \frac{1}{rank_i}\right)$$

where

rank$_i$ = 1,...,α (1 if the first answer is relevant/correct, 2 if the second answer is relevant/correct,...) and = ∞ if none of the α first answers is correct

n = number of queries

Evaluation measures: information retrieval

- When too many documents to judge manually: principle of **depth pooling**:
  - The union of the top *k* documents retrieved by each system corresponding to a given query is built
  - The documents in this depth *k* pool are judged for relevancy with respect to the query

Evaluation measures: subtasks

- E.g., evaluating text categorization (= assignment of controlled language index terms), information extraction, text summarization, …

- **Intrinsic evaluation**:
  - comparing the answers of the system to the answers of the expert
- **Extrinsic evaluation**:
  - judges the quality of the task based on how it affects the completion of some other task (e.g., how a summary affects a retrieval task)

Evaluation measures: classification

- Recall, precision, F-measure, accuracy
- Confusion matrix
- ROC curve

Confusion matrix

- Column: gives number of instances classified by system in the specific class
- Row: gives number of instances classified by expert in the specific class
- Easy to see if system confuses two classes
- Built for binary and multi-class classification problems

Confusion matrix

- Confusion matrix of binary classification decisions (e.g., for intrinsic evaluation of e.g., classification in relevant - non-relevant documents):

	System says yes	System says no
Expert says yes	*tp*	*fn*
Expert says no	*fp*	*tn*

where

tp = true positives   *fn* = *false negatives*

fp = false positives   *tn* = *true negatives*

Confusion matrix

recall = $tp / (tp + fn)$

precision = $tp / (tp + fp)$

error rate = $(fp + fn) / (tp + fp + fn + tn)$

accuracy = $(tp + tn) / (tp + fp + fn + tn)$

recall and precision can be combined into F-measure

- **macro-averaging** : the results of the above measures for each class are averaged over classes
- **micro-averaging**: the results of the above measures are averaged over all binary classification decisions

	System says yes	System says no	
Expert says yes	10	10	Class 1
Expert says no	10	970	

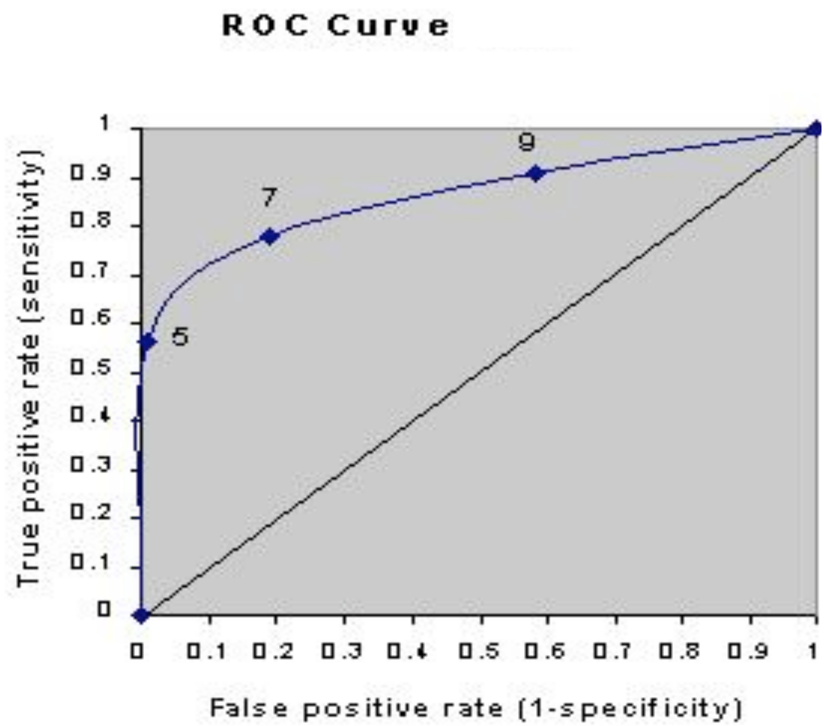
	System says yes	System says no	
Expert says yes	90	10	Class 2
Expert says no	10	890	

	System says yes	System says no	
Expert says yes	100	20	All classification decisions
Expert says no	20	1860	

Macro-averaged precision: $(0.5 + 0.9)/2 = 0.7$
Micro-averaged precision: $100/120 = 0.83$

ROC curve

- **Receiver Operating Characteristic curve**: area under curve should be maximized



1-specificity (= *fp*/(*fp*+*tn*))
sensitivity (= *tp*/(*tp* +*fn*))

Inter-annotator agreement

- **Kappa statistic**: agreement rate when creating 'gold standard' or 'ground truth' corrected for the rate of chance agreement

$$Kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

  where

  $P(A)$= proportion of the annotations on which the annotators agree

  $P(E)$ = proportion of the annotations on which annotations would agree by chance

- *Kappa* > 0.8: good agreement
- 0.67 <= *Kappa* <=0.8: fair agreement
- More than 2 judges: compute average pairwise *Kappa*

Evaluation of summarization and question answering

- **Pyramid method:** comparison of machine-made summary with human-made (model) summaries
  - Human summarizers make summaries with only partially overlapping content
  - From the model summaries a **pyramid of SCUs** (Summary Content Units) is built
  - Machine-made summary is compared with the pyramid and its overlap is measured

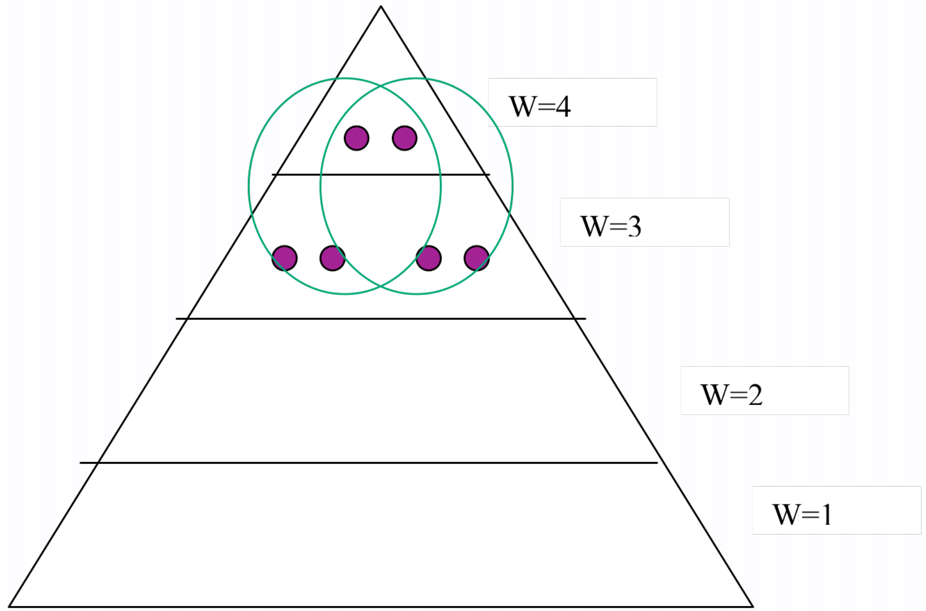
Pyramid method

- **SCU** = collection of paraphrases (words and phrases) that express similar content, e.g.,

  SCU 77 (W=4): Wales has about 3 dozen district councils
  C1: 37 districts in Wales
  C2: 37 district councils
  C3: 38 Welsh districts
  C4: 37 district councils

- Few SCUs are shared by many models, more by only a few models or occur only in 1 model => pyramid of SCUs
- Pyramid of order *n* (*n* = number of tiers of the pyramid, ≤ number of model summaries):
  - we can predict the optimal summary content: it should contain all the SCUs from the top tier, if length permits, also from lower tiers

Pyramid method



Pyramid of order 4: two of six optimal summaries with 4 SCUs

Pyramid method

- For the machine-made summary with *x* SCUs, we compute its summary score *Ss* and the maximally possible summary score *MaxSs* based on the pyramid
- Summary score *Ss*:

$$Ss = \sum_{i=1}^{n} w_i N_i$$

where w_i = weight of the SCUs in tier T_i (possibly chosen as the level number of the pyramid, where T_n is on top of the pyramid, T_1 on the bottom)

N_i = number of SCUs in the summary that appear in T_i

Pyramid method

- The optimal content score *MaxSs* for a summary with *x SCUs* is:

$$MaxSs = \sum_{i=j+1}^{n} w_i |T_i| + w_j (x - \sum_{i=j+1}^{n} |T_i|)$$

  where

$$j = \max_i \left(\sum_{t=i}^{n} |T_t| \geq x \right)$$

- Pyramid score of machine-made summary: $\dfrac{Ss}{MaxSs}$

Evaluation measures: subtasks

- Many of the subtasks in text retrieval regard the processing of texts:

  a number of other criteria for system performance become important :
  - the linguistic coverage: the types of linguistic phenomena a system is able to process
  - the domain coverage
  - the extensibility: the possibility to enlarge the linguistic and/ or domain coverage
  - the portability: the capacity of a system to be transferred from a language and/or domain to another one without major modifications

Evaluation measures: subtasks

- time to train the system
- cost of annotating the training corpus, when such a corpus is needed
- the robustness: the capacity of a system to produce acceptable results even in case of a partial coverage of the linguistic phenomena or the domain
- the linguistic quality of the input and output (e.g., coherence and grammaticality of a summary)
- the granularity of the textual unit processed (whole document, chapter, sentence, …)
- the possibility for the end user to modify some parameters,
- …

Test collections

- **Text retrieval**:
  - TREC: Text REtrieval Conference: http://trec.nist.gov/
    and Text Analysis Conference (TAC):
    http://www.nist.gov/tac/
  - CLEF: Cross Language Evaluation Forum: http://www.clef-campaign.org/
    - Also ImageCLEF: cross-language cross-media retrieval
  - NTCIR: (NII Test Collection for IR Systems) Project: http://research.nii.ac.jp/ntcir/
  - INEX: Initiative for the Evaluation of XML retrieval: http://www.inex.otago.ac.nz/

Test collections

- **Text categorization:**
  - Reuters collection: e.g., Reuters-21578

  http://www.daviddlewis.com/resources/testcollections/

  e.g., Reuters RCV1 (810,000 stories)

  http://about.reuters.com/researchandstandards/corpus/statistics/index.asp
- **Information extraction**:
  - MUC: Message Understanding Conference

    ACE: Automatic Content Extraction: http://www.nist.gov/speech/tests/ace/

Test collections

- **Text summarization**:
  - DUC: Document Understanding Conference: http://duc.nist.gov/
- **Text summarization, textual entailment and question answering:**
  - TAC: Text Analysis Conference: http://www.nist.gov/tac/
- **Collection of queries**:
  - America On Line: http://fack.org/AOL-user-ct-collection/

What have we learned?

- Most common evaluation metrics used in:
  - Information retrieval
  - Text classification and information extraction
  - Text summarization
- Major test collections and competitions

Research questions to be solved

- Metrics that effectively deal with graded relevance judgments instead of binary relevant non-relevant decisions
- Metrics for measuring the performance of complex (e.g., pipelined) tasks
- Metrics that take into account the severeness of errors in a certain context

Further reading

Croft, W.B, Metzler, D. & Strohman, T. (2009). *Search Engines: Information Retrieval in Practice*. Addison Wesley (Chapter 8).

Lewis, D., Yang, Y., Rose, T. & Li, F. (2004). A new benchmark collection for text categorization research. *Journal of Machine Learning* Research, 6, 341-361.

Nenkova, A., Passonneau, R.J. & McKeown, K. (2007). The Pyramid method: Incorporating human content selection variation in summarization evaluation. In *ACM Transactions on Speech and Language Processing,* 4(2),

Voorhees, E.M. & Harman, D.K. (Eds.) (2005). *TREC Experiment and Evaluation in Information Retrieval*. Cambridge, MA: The MIT Press (Chapters1, 2 and 3).