Processing of large document collections

Part 3 (Evaluation of text classifiers, applications of text categorization)

Helena Ahonen-Myka

Spring 2005

Evaluation of text classifiers

- evaluation of document classifiers is typically conducted experimentally, rather than analytically
- reason: in order to evaluate a system analytically, we would need a formal specification of the problem that the system is trying to solve
- text categorization is non-formalisable

Evaluation

- the experimental evaluation of a classifier usually measures its effectiveness (rather than its efficiency)
  - effectiveness= ability to take the right classification decisions
  - efficiency= time and space requirements

Evaluation

- after a classifier is constructed using a training set, the effectiveness is evaluated using a test set
- the following counts are computed for each category i:
  - TP_i: true positives
  - FP_i: false positives
  - TN_i: true negatives
  - FN_i: false negatives

Evaluation

- TP_i: true positives w.r.t. category c_i
  - the set of documents that both the classifier and the previous judgments (as recorded in the test set) classify under c_i
- FP_i: false positives w.r.t. category c_i
  - the set of documents that the classifier classifies under c_i, but the test set indicates that they do not belong to c_i

Evaluation

- TN_i: true negatives w.r.t. c_i
  - both the classifier and the test set agree that the documents in TN_i do not belong to c_i
- FN_i: false negatives w.r.t. c_i
  - the classifier do not classify the documents in FN_i under c_i, but the test set indicates that they should be classified under c_i

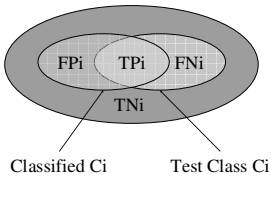
Evaluation measures

- Precision wrt c_i

$$\pi_i = \frac{TP_i}{TP_i + FP_i}$$

- Recall wrt c_i

$$\rho_i = \frac{TP_i}{TP_i + FN_i}$$



7

Evaluation measures

- for obtaining estimates for precision and recall in the collection as a whole, two different methods may be adopted:
  - microaveraging
    - counts for true positives, false positives and false negatives for all categories are first summed up
    - precision and recall are calculated using the global values
  - macroaveraging
    - average of precision (recall) for individual categories

8

Evaluation measures

- microaveraging and macroaveraging may give quite different results, if the different categories have very different generality
- e.g. the ability of a classifier to behave well also on categories with low generality (i.e. categories with few positive training instances) will be emphasized by macroaveraging
- choice depends on the application

9

Combined effectiveness measures

- neither precision nor recall makes sense in isolation of each other
- the trivial acceptor (each document is classified under each category) has a recall = 1
  - in this case, precision would usually be very low
- higher levels of precision may be obtained at the price of lower values of recall

10

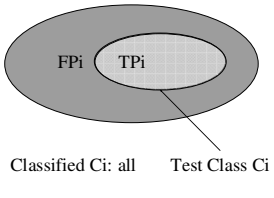
Trivial acceptor

- Precision wrt c_i

$$\pi_i = \frac{TP_i}{TP_i + FP_i}$$

- Recall wrt c_i

$$\rho_i = \frac{TP_i}{TP_i + FN_i}$$



11

Combined effectiveness measures

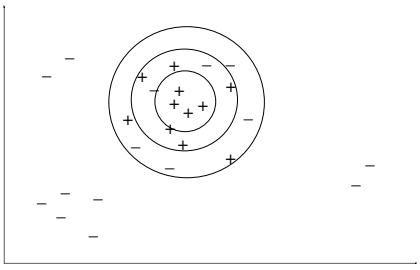
- a classifier should be evaluated by means of a measure which combines recall and precision
- some combined measures:
  - 11-point average precision
  - the breakeven point
  - F1 measure

12

11-point average measure

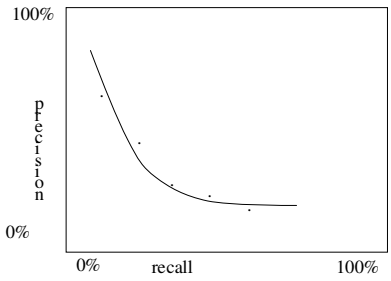
- in constructing the classifier, the threshold is repeatedly tuned so as to allow recall (for the category) to take up values 0.0, 0.1., ..., 0.9, 1.0.
- precision (for the category) is computed for these 11 different values of precision, and averaged over the 11 resulting values

13



14

Recall-precision curve



15

Breakeven point

- process analoguous to the one used for 11-point average precision
  - precision as a function of recall is computed by repeatedly varying the thresholds
- breakeven is the value where precision equals recall

16

F_1 measure

- F_1 measure is defined as:

$$F_1 = \frac{2\pi\rho}{\pi + \rho}$$

- the breakeven point of a classifier is always less or equal than its F_1 value
- for the trivial acceptor, $\pi \to 0$ and $\rho = 1$, $F_1 \to 0$

17

Effectiveness

- once an effectiveness measure is chosen, a classifier can be tuned (e.g. thresholds and other parameters can be set) so that the resulting effectiveness is the best achievable by that classifier

18

Evaluation measures

- efficiency (= time and space requirements)
  - seldom used, although important for real-life applications
  - difficult: environment parameters change
  - two parts
    - training efficiency – average time it takes to build a classifier for a category from a training set
    - classification efficiency – average time it takes to classify a new document under a category

19

Conducting experiments

- in general, different sets of experiments may be used for cross-classifier comparison only if the experiments have been performed
  - on exactly the same collection (same documents and same categories)
  - with the same split between training set and test set
  - with the same evaluation measure

20

Applications of text categorization

- automatic indexing for Boolean information retrieval systems
- document organization
- text filtering
- word sense disambiguation
- authorship attribution
- hierarchical categorization of Web pages

21

Automatic indexing for information retrieval systems

- in an information retrieval system, each document is assigned one or more keywords or keyphrases describing its content
  - keywords belong to a finite set called controlled dictionary
- TC problem: the entries in a controlled dictionary are viewed as categories
  - $k_1 \leq x \leq k_2$ keywords are assigned to each document

22

Document organization

- indexing with a controlled vocabulary is an instance of the general problem of document collection organization
- e.g. a newspaper office has to classify the incoming "classified" ads under categories such as Personals, Cars for Sale, Real Estate etc.
- organization of patents, filing of newspaper articles...

23

Text filtering

- classifying a stream of incoming documents by an information producer to an information consumer
- e.g. newsfeed
  - producer: news agency; consumer: newspaper
  - the filtering system should block the delivery of documents the consumer is likely not interested in

24

Word sense disambiguation

- given the occurrence in a text of an ambiguous word, find the sense of this particular word occurrence
- e.g.
  - bank, sense 1, like in ”Bank of Finland”
  - bank, sense 2, like in ”the bank of river Thames”
  - occurrence: ”Last week I borrowed some money from the bank.”

25

Word sense disambiguation

- indexing by word senses rather than by words
- text categorization
  - documents: word occurrence contexts
  - categories: word senses
- also resolving other natural language ambiguities
  - context-sensitive spelling correction, part of speech tagging, prepositional phrase attachment, word choice selection in machine translation

26

Authorship attribution

- task: given a text, determine its author
- author of a text may be unknown or disputed, but some possible candidates and samples of their works exist
- literary and forensic applications
  - who wrote this sonnet? (literary interest)
  - who sended this anonymous letter? (forensics)

27

Hierarchical categorization of Web pages

- e.g. Yahoo like web hierarchical catalogues
- typically, each category should be populated by ”a few” documents
- new categories are added, obsolete ones removed
- usage of link structure in classification
- usage of the hierarchical structure

28