Evaluating Text Categorization[1]

Appeared (with same pagination) in *Proceedings of the Speech and Natural Language Workshop*, Asilomar, Feb. 1991. Morgan Kaufmann, San Mateo, CA, pp. 312–318.

David D. Lewis

Computer and Information Science Dept.
University of Massachusetts
Amherst, MA 01003

ABSTRACT

While certain standard procedures are widely used for evaluating text retrieval systems and algorithms, the same is not true for text categorization. Omission of important data from reports is common and methods of measuring effectiveness vary widely. This has made judging the relative merits of techniques for text categorization difficult and has disguised important research issues.

In this paper I discuss a variety of ways of evaluating the effectiveness of text categorization systems, drawing both on reported categorization experiments and on methods used in evaluating query-driven retrieval. I also consider the extent to which the same evaluation methods may be used with systems for text extraction, a more complex task. In evaluating either kind of system, the purpose for which the output is to be used is crucial in choosing appropriate evaluation methods.

INTRODUCTION

Text classification systems, i.e. systems which can make distinctions between meaningful classes of texts, have been widely studied in information retrieval and natural language processing. The majority of information retrieval research has been devoted to a particular form of text classification—*text retrieval*. Text retrieval systems find or route texts in response to arbitrary user queries or interest profiles. Evaluation has been a focus of research in text retrieval since the beginning, and standard evaluation methods are in wide use.

A smaller, but significant, body of work has examined a task variously known as machine-aided indexing, automated indexing, authority control, or *text categorization*. Text categorization is the assignment of texts to one or more of a pre-existing set of categories, rather than classifying them in response to an arbitrary query. Categorization may be performed for a wide range of reasons, either as an end in itself or as a component of a larger system.

The literature on text categorization is widely scattered and shows little agreement on evaluation methods. This makes it very difficult to draw conclusions about the relative effectiveness of techniques so that, unlike the situation in query-driven retrieval, there is no consensus on a set of basic evaluation methods for text categorization.

In this paper I discuss measures of *effectiveness* for text categorization systems and algorithms. Effectiveness refers to the ability of a categorization to supply information to a system or user that wants to access the texts. Measuring effectiveness is just one of several kinds of evaluation that should be considered [Spa81a, CH88, PF90].

After considering effectiveness evaluation for text categorization we will turn to a related task, text extraction, and consider what role the effectiveness measures discussed for categorization have there. A common theme is the need to consider in an evaluation the purpose for which information is generated from the text.

I will have occasion in the following to repeatedly refer to a chapter by Tague [Tag81] in Sparck Jones' collection on information retrieval experimentation [Spa81a]. This collection discusses a wide range of evaluation issues, and is an important resource for anyone interested in the evaluation of text-based systems.

EFFECTIVENESS MEASURES

While a number of different effectiveness measures have been used in evaluating text categorization in the past, almost all have been based on the same model of decision making by the categorization system. I begin by discussing this *contingency table* model, which motivates a small number of simple and widely used effectiveness measures. Complexities arise, however, in how to compute and interpret these measures in the context of a text categorization experiment. The bulk of the discussion concerns these complexities.

[1] Current Address: Center for Information and Language Studies; University of Chicago; Chicago, IL 60637; *lewis@tira.uchicago.edu*

	Yes is Correct	No is Correct	
Decides Yes	a	b	$a+b$
Decides No	c	d	$c+d$
	$a+c$	$b+d$	$a+b+c+d=n$

Table 1: Contingency Table for a Set of Binary Decisions

The Contingency Table

Consider a system that is required to make n binary decisions, each of which has exactly one correct answer (either Yes or No). The result of n such decisions can be summarized in a contingency table, as shown in Table 1. Each entry in the table specifies the number of decisions of the specified type. For instance, a is the number of times the system decided Yes, and Yes was in fact the correct answer.

Given the contingency table, three important measures of the system's effectiveness are:

(1) recall $= a/(a+c)$

(2) precision $= a/(a+b)$

(3) fallout $= b/(b+d)$

Measures equivalent to recall and fallout made their first appearance in signal detection theory [Swe64], where they play a central role. Recall and precision are ubiquitous in information retrieval, where they measure the proportion of relevant documents retrieved and the proportion of retrieved documents which are relevant, respectively. Fallout measures the proportion of *non*relevant documents which are retrieved, and has also seen considerable use.

A decision maker can achieve very high recall by rarely deciding No, or very high precision (and low fallout) by rarely deciding Yes. For this reason either recall and precision, or recall and fallout, are necessary to ensure a nontrivial evaluation of a decision maker's effectiveness under the above model.

Another measure sometimes used in categorization experiments is overlap:

(4) overlap $= a/(a+b+c)$

This measure is symmetric with respect to b and c, and so is sometimes used to measure how much two categorizations are alike without defining one or the other to be correct.

It is appropriate at this point to mention some of the limitations of the contingency table model. It does not take into account the possibility that different errors have different costs; doing so requires more general decision theoretic models. The contingency table also requires all decisions to be binary. It may be desirable for category assignments to be weighted rather than binary, and we will discuss later one approach to evaluation in this case.

Defining Decisions and Averaging Effectiveness

The contingency table model presented above is applicable to a wide range of decision making situations. In this section, I will first consider how query-driven text retrieval has been evaluated under this model, and then consider how text categorization can be evaluated under the same model. In both cases it will be necessary to interpret the system's behavior as a set of binary decisions.

In a query-driven retrieval systems, the basic decision is whether or not to retrieve a particular document for a particular query. For a set of q queries and d documents a total of $n = qd$ decisions are made. Given those qd decisions, two ways of computing effectiveness are available. *Microaveraging* considers all qd decisions as a single group and computes recall, precision, fallout, or overlap as defined above. *Macroaveraging* computes these effectiveness measures separately for the set of d documents associated with each query, and then computes the mean of the resulting q effectiveness values.

Macroaveraging has been favored in evaluating query-driven retrieval, partly because it gives equal weight to each user query. A microaveraged recall measurement, for instance, would be disproportionately affected by recall performance on queries from users who desired large numbers of documents.

An obvious analogy exists between categories in a text categorization system and queries in a text retrieval system. The most common view taken of categorization is that an assignment decision is made for each category/document pair. A categorization experiment will compare the categorization decisions made by a computer system with some standard of correctness, usually human category assignment. In contrast to evaluations of query-driven retrieval, evaluations of categorization have usually used microaveraging rather than macroaveraging. Many ad hoc variants of both forms of averaging have also been used.

Whether microaveraging or macroaveraging is more informative depends on the purpose for the categorization. For instance, if categorization is used to index documents for text retrieval, and each category appears in user queries at about the same frequency it appears in documents, then

Category Set	Correct	Assigned	R	P	F
ABCD	AB	A C	50	50	50
ABCDEFGH	AB	A C	50	50	16
ABCDEFGHIJKL	AB	A C	50	50	10

Table 2: Recall (R), Precision (P), and Fallout (F) of Categorizer X on One Document

microaveraging seems very appropriate. On the other hand, if categorization were used to route documents to divisions of a company, with each division viewed as being equally important, then macroaveraging would be more informative. The choice will often not be clearcut. I assume microaveraging in the following discussion unless otherwise mentioned.

Precision Versus Fallout

Precision and fallout both measure (in roughly inverse ways) the tendency of the categorizer to assign incorrect categories. However, in doing so they capture different properties of the categorization.

In the context of query-driven retrieval, Salton has pointed out how systems which maintain constant precision react differently to increasing numbers of documents than those which maintain constant fallout [Sal72]. Similar effects can arise for categorizers as the number or nature of categories changes.

Table 2 shows the hypothetical performance of categorizer X as the category set is expanded to include new topics. Decreasing fallout suggests that the categorizer X incorrectly assigns categories in proportion to the number of correct categories to be assigned. A different categorizer, Y, might show the pattern in Table 3, suggesting categories are incorrectly assigned in proportion to the total number of incorrect categories (or in proportion to the total number of all categories).

In extreme cases a system could actually improve on precision while worsening on fallout, or vice versa. Having both measures, plus recall, available is useful in quickly appraising a method's behavior under changing circumstances.

Partitioning of Results

The basic tools of microaveraging and macroaveraging can be applied to arbitrary subsets of categorization decisions. Subsets of decisions can be defined in terms of subsets of categories, subsets of documents, or gradations in the correctness standard.

Category Set	Correct	Assigned	R	P	F
ABCD	AB	A C	50	50	50
ABCDEFGH	AB	A CEF	50	25	50
ABCDEFGHIJKL	AB	A CEFIJK	50	16	50

Table 3: Recall (R), Precision (P), and Fallout (F) of Categorizer Y on One Document

Categories can be partitioned by importance, frequency, similarity of meaning, or strategy used in assigning them. Presenting effectiveness measures averaged over category groups defined by frequency in the training set would be extremely informative, but does not appear to have been done in any published study. If the number of categories is small enough, effectiveness can be presented separately for each category [HKC88].

Subsets of the set of test documents can be defined as well, particularly if the behavior of the system on texts of different kinds is of interest. Maron grouped documents on the basis of the amount of evidence they provided for making a categorization decision, and showed that effectiveness increased in proportion to the amount of evidence [Mar61].

Finally, it is sometimes appropriate to partition results by degree of correctness of a category/document pair. While the contingency table model assumes that an assignment decision is either correct or incorrect, the standard they are being tested against may actually have gradations of correctness. The model can still be used if gradations are partitioned into two disjoint classes, for instance *correct* and *marginal* being considered correct, and *ineffective* and *incorrect* being considered incorrect. In this circumstance, it may be desirable to present results under several plausible partitions.

The appropriate partitions to make will depend on many factors that cannot be anticipated here. A crucial point to stress, however, is that care should be taken to partition supporting data on the task and system in the same fashion [Lew91]. For instance, if effectiveness measures are presented for subsets of documents, then statistics such as average number of words per document, etc. should be given for the same groups of documents.

Arithmetic Anomalies

The above discussion assumed that computing the effectiveness measures is always straightforward. Referring to equations (1) to (3) shows that 0 denominators arise when there exist no correct category assignments, no incorrect category assignments, or when the system never assigns a category. All these situations are extremely unlikely when microaveraging is used, but are quite possible under macroaveraging.

For evaluating query-driven retrieval, Tague suggests either treating 0/0 as 1.0 or throwing out the query, but says neither solution is entirely satisfactory. For a categorization system, we also have the option of partitioning the categories and macroaveraging only over the categories for which these anomalies don't arise. As discussed above, the

same partitioning should be used for any background data presented on the testset and task.

One Category or Many?

Evaluations of systems which assign multiple categories to a document have often been flawed, particularly for categorizers which use statistical techniques. For instance, some of the results in [Mar61] and [KW75] were obtained under assumptions equivalent to the categorizer knowing in advance how many correct categories each test document has. This knowledge is not available in an operational setting.

Better attempts to both produce and evaluate multiple category assignments are found in work by Fuhr and Knorz, and by Field. Field uses the strategy of assigning the top k categories to a document, but unlike the above studies does this without knowledge of the proper number of categories for any particular document. He then plots the recall value achieved for variations in the number of categories assigned [Fie75]. Fuhr and Knorz plot a curve showing tradeoff between recall and precision as a category assignment threshold varies [FK84].

When categories are completely disjoint and a categorizer always assigns exactly 1 of the M categories to a text, we really have a single M-ary decision, rather than M binary decisions. The contingency table model provides one way of summarizing M-ary decision effectiveness, but other approaches, such as confusion matrices [Swe64], may be more revealing.

Standard of Correctness

The effectiveness measures described above require that correct categorizations are known for a set of test documents. In cases where an automated categorizer is being developed to replace or aid manual categorization, categorizations from the operational human system may be used as the standard. Otherwise, it may be necessary to have human indexers categorize some texts specifically for the purposes of the experiment.

Many studies have found that even professional bibliographic indexers disagree on a substantial proportion of categorization decisions [Bor64, Fie75, HZ80]. This calls into question the validity of human category assignment as a standard against which to judge mechanical assignment. One approach to this problem has been to have an especially careful indexing done [Fie75, HZ80]. Sometimes evaluation is done against several indexings [Fie75, HKC88].

Another approach is to accept that there will always be some degree of inconsistency in human categorization, and that this imposes an upper limit on the effectiveness of machine categorization. The degree of consistency between several human indexers can be measured, typically using overlap, as defined in Equation (4), or some variant of this.

How measures of consistency between human indexers might best aid the interpretation of machine categorization effectiveness is unclear. Overlap between the machine-assigned categories and each human indexers' categories can be measured and compared to overlap among humans. It is less clear how to interpret recall, precision, or fallout in the presence of a known level of inconsistency.

The possibility also exists that machine categorization could be *better* than human categorization, making consistency with human categorization a questionable measure under any circumstance. Indirect evaluation, discussed in the next section, is the best way to address this possibility.

Indirect Evaluation

The output of a text categorization system is often used by another system in performing text retrieval, text extraction, or some other task. When this is the case, it is possible to evaluate the categorization indirectly, by measuring the performance of the system which uses the categorization. This indirect evaluation of the categorization can be an important complement to direct evaluation, particularly when multiple categorizations are available to be compared.

How an indirect evaluation is done depends on the kind of system using the categorized text. Most categorizers have been intended to index documents for query-driven text retrieval. Despite this, there have been surprisingly few studies [Har82, FK84] comparing text retrieval performance under different automatic category assignments.

The focus on manual categorization as a standard appears to have led categorization researchers to ignore some promising research directions. For instance, I know of no study that has evaluated *weighted* assignment of categories to documents, despite early recognition of the potential of this technique [Mar61] and the strong evidence that weighting free text terms in documents improves retrieval performance [Sal86].

Categorization of documents may be desired for other purposes than supporting query-driven retrieval. Separation of a text stream by category may allowing packaging of the text stream as different products [Hay90]. Some comparison of average retrieval effectiveness across text streams might be an appropriate measure in this case.

Categorization may also be used to select a subset of texts for more sophisticated processing, such as extraction of information or question answering [JR90]. Evaluating the quality of the extracted information may give some insight into categorization performance, though the connection can be distant here.

There are drawbacks to indirect evaluation, of course. Tague questions why any particular set of queries should serve as a test of an indexing. Clearly, if a categorization is to be evaluated by text retrieval performance, the query set needs to be as large as possible, and representative of the actual usage the system will experience. When categorization is used as a component in a complex language understanding system, that system itself may be difficult to evaluate [JR90] or differences in categorization quality may be hard to discern from overall system behavior. A single categorization may also be intended to serve several purposes, some possibly not yet defined. Using both direct and indirect evaluation will be the best approach, when practical.

Other Issues

The evaluation of natural language processing (NLP) systems is an area of active research [PF90], and a great deal remains to be learned. Much more could be said even about evaluating categorization systems. In particular, I have focused entirely on numerical measures. Carefully chosen examples, examined in detail, can also be quite revealing [HKC88]. However, the numerical measures described above provide a useful standard for understanding the differences between methods under a variety of conditions.

Comparison between categorization methods would be aided by the use of common testsets, something which has rarely been done. (An exception is [BB64].) Development of standard collections would be an important first step to better understanding of text categorization.

Categorization is an important facet of many kinds of text processing systems. The effectiveness measures defined above may be useful for evaluating some aspects of these systems. In the next section we consider the evaluation of text extraction systems from this standpoint.

IMPLICATIONS FOR EVALUATING TEXT EXTRACTION

Systems for *text extraction* generate formatted data from natural language text. Some forms of extraction, for instance specifying the highest level of action in a naval report [Sun89], are in fact categorization decisions. Other forms of extraction are very different, and do not fit well into the contingency table model.

In the following I briefly consider evaluation of text extraction systems using the effectiveness measures described for categorization. Two perspectives are taken—one focusing on the type of data extracted and the other focusing on the purpose for which extraction is done.

Types of Extracted Data

Extracted data can include binary or M-ary categorizations, quantities, and templates or database records with atomic or structured fillers [Sun89, McC90, Hal90]. The number of desired records per text may depend on text content, and cross references between fillers of record fields may be required.

Using the effectiveness measures described above requires interpreting the system output in terms of a set of binary decisions which can be either correct or incorrect. The measures become less meaningful as the extraction task becomes less a matter of making isolated decisions with easily defined correctness, and more a matter of generating a legal expression from some potentially infinite language.

Binary data, either as the sole output of extraction or as the filler of a fixed subpart of a larger structure, fits easily into the contingency table model of evaluation. This includes the case where a slot can have 0 or more fillers from a fixed set of possible fillers. Each pair of the form (slot, possible filler) can be treated as a category in the categorization model. Micro- or macroaveraging across slot/filler pairs for a single slot or for all slots in a template can be done. The situation where exactly one of a fixed set of M fillers must fill a slot is an M-ary decision, as mentioned above for categorization.

Another common extraction task is to recognize all human names in a piece of text, and produce a *canonical string* for each name as part of the extracted data. Effectiveness measures from categorization begin to break down here. Treating the assignment of each possible canonical name as a binary decision is likely to be uninformative, given the very large set of legal names. (And is impossible if instead of a fixed set of canonical names there are rules defining an unbounded number of them.) The situation is even more difficult when arbitrary strings may be slot fillers.

The MUC-3 evaluation [Hal90] has taken the approach of retaining the contingency table measures but redefining the set of possible decisions. Rather than taking the cross-product of the set of all fillers and the set of all documents, the set of decisions is implicitly considered to be the union of all correct string/document assignments and all system-produced string/document assignments. This is equivalent to setting cell d of the contingency table to 0, while retaining the others. Fallout is thus eliminated as a measure but recall, precision, and overlap can still be computed. A scheme for assigning partial credit is also used.

While this approach has been quite useful, it may not be ideal. Two processes are being evaluated at once—recognition of an extractable concept, and selection of a string (canonical or arbitrary) to represent that concept. It may be preferable, for instance, to evaluate these processes separately. This approach also requires subtle human

judgments of the relative correctness of various strings that might be extracted. Finally, when comparing systems using this approach, the underlying decision spaces may be different for each system, making interpreting the effectiveness measures more difficult.

When a system goes beyond string fills to filling slots with arbitrary structures, the contingency table model becomes very difficult to apply. At best there may be some hopes of capturing some parts of the task in this way, such as getting the right category of structure in a slot. More research on evaluation is clearly needed here.

Purposes for Extracted Data

The data type of extracted information affects what effectiveness measures can be computed. Even more important, however, is the purpose for which information is being extracted. This issue has been given surprisingly little attention in published discussions of text extraction systems. In the following, I give three examples to suggest that explicit consideration of how extracted data will be used is crucial in choosing appropriate effectiveness measures.

Statistical Analysis of Real-World Events A database of extracted information may be meant to support queries about real-world events described in the texts. An analyst might want to check for correlations between numbers of naval equipment failures and servicing in certain ports, or list the countries where plastic explosives have been used in terrorist bombings, to give examples.

Accurate answers to questions about numbers of events depend on recognizing when multiple event references in the same or in different documents in fact refer to a single real world event, and on proper handling of phenomena such as plurals, numbers, and quantification. High precision and low fallout may be favored over high recall. If it is expected that the same event will be described by multiple sources, a single failure to recognize it may not be important. Evaluation might focus on effectiveness in extracting details necessary to uniquely identify each event. On the other hand, if support of arbitrary existence queries (*Has plastic explosive been used...*) is important, then recall for all recognizable details of events may be the most important thing to evaluate.

The degree of connection between reports of events and actual events will vary from reliable (intra-agency traffic) to dubious (political propaganda). This makes it likely that the extraction system will at best be an aid to a human analyst, who will need to make judgment calls on the reliability of textual descriptions. The most useful evaluation may be of the analyst's performance with and without the extraction system.

Content Analysis *Content analysis* has been defined in many different ways ([Hol69], pp. 2–3) but here I focus particularly on the analysis of texts to gain insight into the motivations and plans of the texts' authors. In its simplest form content analysis involves counting the number of occurrences of members of particular linguistic classes. For instance, one might count how often words with positive or negative connotations are used in referring to a neighboring country. The great potential of the computer to aid with the drudgery of analyzing large corpora of text has long been recognized, as has the potential for NLP to improve the effectiveness of this process.

In content analysis, faithfulness to the text rather than faithfulness to the world may be the primary concern. Of particular importance is that the number of instances of a particular linguistic item extracted is not affected by extraneous variables. Consider a comparison of the number of references to a particular border skirmish in political broadcasts from two countries. In this case, one would want confidence that extraction effectiveness was about the same for texts from the two countries and was not affected by, for instance, differing capitalization conventions. The absolute level of effectiveness might be a lesser concern.

Indexing for Query-Driven Text Retrieval In this case, the extracted data is used only indirectly. An analyst will use either a text retrieval system or a conventional database system to retrieve documents indexed by extracted data. The analyst may want the documents for any of a number of purposes, including the ones described above. The difference is that extracted information participates in the analysis only to the extent of influencing which documents the analyst sees. No numeric values are derived directly from the extracted data.

In evaluating formatted data extracted for this purpose, a number of results from information retrieval research are important to consider. One is the fact, mentioned earlier, that document-specific weighting of indexing units is likely to substantially increase performance. Since NLP systems can potentially use many sources of evidence in deciding whether to extract a particular piece of information, there is a rich opportunity for such weighting.

Another lesson from IR research is that people find it very difficult to judge the quality of indexing in the absence of retrieval data. Strongly held intuitions about the relative effectiveness of indexing languages and indexing methods for supporting document retrieval have often been shown by experiment to be incorrect [Spa81b]. If the primary purpose of extracted information is to support querying, then indirect evaluation, i.e. testing with actual queries, is very important.

CONCLUSION

Text categorization plays a variety of roles in text-based systems. Evaluation of categorization effectiveness is important, both for confidence in operational systems and for progress in research. Several good measures, based on a model of binary decision making, are available for evaluating the effectiveness of a categorization. I have discussed some of the issues to consider in using these measures, and stressed that the purpose for which categorization is being done needs to be considered. The use of both indirect and direct evaluation is preferable. I also discussed how some of the work done by text extraction systems can be viewed as categorization and evaluated in a similar fashion, though new measures are needed as well.

Acknowledgments

The author is preparing a longer article on this topic [Lew91] so comments on the above would be most welcome. This research has already greatly benefited from discussions with Laura Balcom, Nancy Chinchor, Bruce Croft, Ralph Grishman, Jerry Hobbs, Adele Howe, Lisa Rau, Penelope Sibun, Beth Sundheim, and Carl Weir. The research has been supported by AFOSR under grant AFOSR-90-0110, by the NSF under grant IRI-8814790, and by an NSF Graduate Fellowship.

REFERENCES

[BB64] Harold Borko and Myrna Bernick. Automatic document classification part II. additional experiments. *J. Association for Computing Machinery*, 11(2):138–151, April 1964.

[Bor64] Harold Borko. Measuring the reliability of subject classification by men and machines. *American Documentation*, pages 268–273, October 1964.

[CH88] Paul R. Cohen and Adele E. Howe. How evaluation guides AI research. *AI Magazine*, pages 35–43, Winter 1988.

[Fie75] B. J. Field. Towards automatic indexing: Automatic assignment of controlled-language indexing and classification from free indexing. *J. Documentation*, 31(4):246–265, December 1975.

[FK84] N. Fuhr and G. E. Knorz. Retrieval test evaluation of a rule based automatic indexing (AIR/PHYS). In C. J. van Rijsbergen, editor, *Research and Development in Information Retrieval*, pages 391–408, Cambridge. Cambridge University Press.

[Hal90] Peter C. Halverson. MUC scoring system: user's manual. Edition 1.5, General Electric Corporate Research and Development, Schenectady, NY, November 2 1990.

[Har82] P. Harding. Automatic Indexing and Classification for Mechanised Information Retrieval. BLRDD Report No. 5723, British Library R & D Department, London, February 1982.

[HKC88] Philip J. Hayes, Laura E. Knecht, and Monica J. Cellio. A news story categorization system. In *Second Conference on Applied Natural Language Processing*, pages 9–17, 1988.

[Hay90] Philip J. Hayes. Intelligent high-volume text processing using shallow, domain-specific techniques. In P. S. Jacobs, editor, *Text-Based Intelligent Systems*, pages 70–74, Schenectady, NY, 1990. GE R & D Center. Report Number 90CRD198.

[Hol69] Ole R. Holsti. *Content Analysis for the Social Sciences and Humanities*. Addison-Wesley, Reading, MA, 1969.

[HZ80] Karen A. Hamill and Antonio Zamora. The use of titles for automatic document classification. *J. American Society for Information Science*, pages 396–402, 1980.

[JR90] Paul S. Jacobs and Lisa F. Rau. SCISOR: Extracting information from on-line news. *Communications of the ACM*, 33(11):88–97, November 1990.

[KW75] B. Gautam Kar and L. J. White. A distance measure for automatic sequential document classification. Technical Report OSU-CISRC-TR-75-7, Computer and Information Science Research Center; Ohio State Univ., Columbus, Ohio, August 1975.

[Lew91] David D. Lewis. Evaluating text classification systems. In preparation., 1991.

[Mar61] M. E. Maron. Automatic indexing: An experimental inquiry. *J. of the Association for Computing Machinery*, 8:404–417, 1961.

[McC90] Rita McCardell. Evaluating natural language generated database records. In *Proceedings of Speech and Natural Language Workshop*, pages 64–70. Defense Advanced Research Projects Agency, Morgan Kaufmann, June 1990.

[PF90] Martha Palmer and Tim Finin. Workshop on the evaluation of natural language processing systems. *Computational Linguistics*, 16:175–181, September 1990.

[Sal72] G. Salton. The "generality" effect and the retrieval evaluation for large collections. *J. American Society for Information Science*, pages 11–22, January–February 1972.

[Sal86] Gerard Salton. Another look at automatic text-retrieval systems. *Communications of the ACM*, 29(7):648–656, July 1986.

[Spa81a] Karen Sparck Jones, editor. *Information Retrieval Experiment*. Butterworths, London, 1981.

[Spa81b] Karen Sparck Jones. Retrieval system tests 1958–1978. In Karen Sparck Jones, editor, *Information Retrieval Experiment*, chapter 12. Butterworths, London, 1981.

[Sun89] Beth M. Sundheim. Plans for task-oriented evaluation of natural language understanding systems. In *Proceedings of the Speech and Natural Language Workshop*, pages 197–202. Defense Advanced Research Projects Agency, Morgan Kaufmann, February 1989.

[Swe64] John A. Swets, editor. *Signal Detection and Recognition by Human Observers*. John Wiley & Sons, New York, 1964.

[Tag81] Jean M. Tague. The pragmatics of information retrieval experimentation. In Karen Sparck Jones, editor, *Information Retrieval Experiment*, chapter 5. Butterworths, London, 1981.