Towards Language Independent Automated Learning of Text Categorization Models

Chidanand Apté
IBM Research Division
T.J. Watson Research Center
Yorktown Heights, NY 10598
apte@watson.ibm.com

Fred Damerau
IBM Research Division
T.J. Watson Research Center
Yorktown Heights, NY 10598
damerau@watson.ibm.com

Sholom M. Weiss
Rutgers University
Dept. of Computer Science
New Brunswick, NJ 08903
weiss@cs.rutgers.edu

Abstract

We describe the results of extensive machine learning experiments on large collections of Reuters' English and German newswires. The goal of these experiments was to automatically discover classification patterns that can be used for assignment of topics to the individual newswires. Our results with the English newswire collection show a very large gain in performance as compared to published benchmarks, while our initial results with the German newswires appear very promising. We present our methodology, which seems to be insensitive to the language of the document collections, and discuss issues related to the differences in results that we have obtained for the two collections.

1 Introduction

In many carefully organized text storage and retrieval systems, texts are classified with one or more codes chosen from a classification system. Examples include the NTIS (National Technical Information Service) documents from the US government, news services like UPI and Reuters, publications like the ACM Computing Reviews and many others. Assigning subject classification codes manually to documents is time consuming and expensive. Recent work has shown that in certain environments, knowledge based systems can do code assignment quickly and accurately [Hayes and Weinstein, 1991, Hayes *et al.*, 1990]. Human-engineered rule-based models for assigning subject codes, while relatively effective, are also very expensive in time and effort for their development and continued support. Machine learning methods provide an interesting alternative for automating the rule construction process. This report presents results on experiments to derive the assignment rules automatically from samples of the text to be classified.

A well known example of a knowledge-based system for the classification task is the CONSTRUE system [Hayes *et al.*, 1990] used by the Reuters news service. This is a rule based expert system using manually constructed rules to assign subject categories to news stories, with a reported recall and precision of over 90% on 750 test cases [Hayes and Weinstein, 1991]. While these are exceptionally good results, the test set seems to have been relatively sparse when compared to the number of possible topics. An example of a machine learning system for the same task is a system based on Memory Based Reasoning [Masand *et al.*, 1992], which employs nearest neighbor style classification and has a reported accuracy in the range of 70-80% on Dow Jones news stories.

In considering the problem of categorizing documents, the rule based approach has considerable appeal. While weighted solutions such as the linear probabilistic methods used in [Lewis, 1992b] or nearest-neighbor methods may also prove reasonable, the models they employ are not explicitly interpretable. Since human-engineered systems have been successfully constructed using rule-based solutions, it would be most useful to continue with a model that is compatible with human-expressed knowledge. Because of the parsimonious and interpretable nature of decision rules, we can readily augment our knowledge or verify the rules by examining related categorized documents.

We report here results that we have obtained with using a rule based machine learning approach on two large collections of Reuters' newswires, in English and German. The collections are essentially streams of stories, numbering in the tens of thousands. Each story has associated with it a headline, a date, one or more topics (that have been assigned by Reuters staff), and various fragments of information mainly used for book-keeping purposes. The goal is to have the computer system induce pattern directed rules that can be used for automatically assigning the topics. Amongst our more interesting observations was the fact that our methodology seems to perform almost equally well for both the English and the German collection, suggesting that our approach may be insensitive to language issues, thereby making it a more versatile and portable technique for document classification.

2 Automated Learning of Topic Assignment Models

Machine learning systems solve problems by examining samples described in terms of measurements or features. For machine learning methods to be applicable, the samples of documents must be transformed into this type of representation. For text categorization, an adaptation of a machine learning method must implement the following main processes:

- A preprocessing step for determining the values of the features or attributes that will used for representing the individual documents within a collection. This is essentially the *dictionary* creation process.
- A representation step for mapping each individual document into a *training sample* using the above dictionary, and associating it with a *label* that identifies its category.
- An induction step for finding patterns that distinguish categories from one another.
- An evaluation step for choosing the *best* solution, based on minimizing the classification error or cost.

The initial task is to produce a list of attributes from samples of text of labeled documents, i.e., the dictionary. The attributes are single words or word phrases. Given an attribute list, sample cases can be described in terms of the words or phrases found in the documents. Each case consists of the values of the attributes for a single article, where the values could be either boolean, i.e., indicating whether the attribute appears in the text or does not, or numerical, i.e., frequency of occurrence in the text being processed. In addition, each case is labeled to indicate the classification or topic of the article it represents.

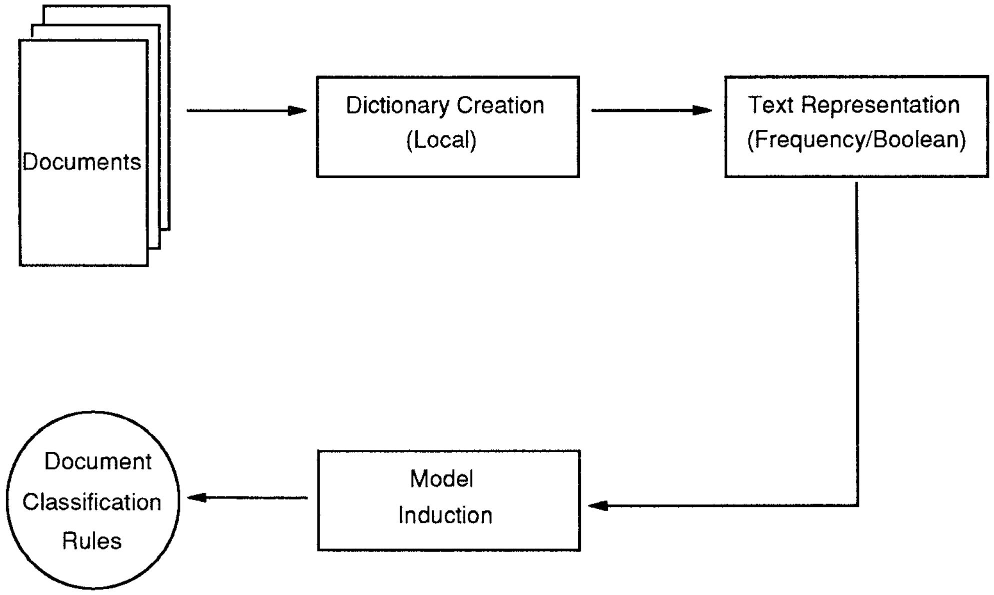

Figure 1: Machine Learning Architecture for Document Classification

For rule induction, the objective is to find sets of decision rules that distinguish one category of text from the others. The best rule set is selected, where "best" is a rule set that is both accurate and not excessively complex. Accuracy of rule sets can be effectively measured on large numbers of independent test cases. Complexity can be measured in terms of numbers of rules or rule components, where smaller rule sets that are reasonably close to the best accuracy are sometimes preferred to more complex rules sets with slightly greater accuracy.

Figure 1 illustrates our strategy. Here a *local* dictionary is created for each classification topic. Only single words found in documents on the given topic are entered in the local dictionary. For each local dictionary, the n most frequently occurring words are used as features. In addition, instead of boolean features, frequency counts of the occurrence of words in a story, or more complicated frequency related measures, can be used.

Research on machine learning from text suggests that the simpler dictionaries of single words give the best performance. This does *not* mean that the best solution ignores phrases and combinations of words. Clearly these combinations are important to understanding text. Rather, the burden is shifted from a preprocessing program that composes a dictionary to a learning program that finds a solution. Thus these research results mostly suggest that it is very difficult to find the right combinations of words independent of the ultimate decision model. The implication of this analysis is that performance can be increased by improved learning methods. These methods should potentially find higher order relationships in the feature space i.e. the dictionary words.

One of the main distinguishing characteristics of our approach is that we will use a rule induction model for our representation. An example of these rules is illustrated in Table 1. Here the problem is posed as a two class problem. If any of these rules is satisfied, the story is classified as "football". Otherwise, the decision reverts to the default class of a "non-football" article. Most applications of text classification involve classes that are not exclusive, and one or more of the categories can occur simultaneously. Thus most problems are handled as multiple two-class problems.

Rule	Class
running back	football article
kicker	football article
injure reserve	football article
award & player	football article

Table 1: Example of an Induced Rule Set

Rule and tree induction methods have been extensively described in published works [Breiman *et al.*, 1984, Weiss and Kulikowski, 1991, Quinlan, 1993]. For our document indexing system, we have used a rule induction technique called Swap-1 [Weiss and Indurkhya, 1993]. Rule induction methods attempt to find a compact "covering" rule set that completely partitions the examples into their correct classes. The covering set is found by heuristically searching for a single best rule that covers cases for only one class. Having found a best conjunctive rule for a class C, the rule is added to the rule set, and the cases satisfying it are removed from further consideration. The process is repeated until no cases remain to be covered. Unlike decision tree induction programs and other rule induction methods, Swap-1 has an advantage in that it uses optimization techniques to revise and improve its decisions. Once a covering set is found that separates the classes, the induced set of rules is further refined by either pruning or statistical techniques. Using train and test evaluation methods, the initial covering rule set is then scaled to back to the most statistically accurate subset of rules.

	TRAINING CASES	
	Football	Not Football
Football	151	10
Not Football	0	1081

	TEST CASES	
	Football	Not Football
Football	135	26
Not Football	12	1069

Table 2: Example of Estimated Performance for a Rule Set

For the document classification application, Swap-1 induces rules that represent patterns, i.e. combinations of attributes, that determine the most likely class for an article. The result of applying Swap-1 to a training set of cases is a set of rules and a table of the associated error rates on the training as well as test samples. The results for applying the rule set of Table 1 are illustrated in Table 2. A detailed discussion of Swap-1 and its use in our document classification experiments appears in [Apté *et al.*, 1993].

3 Results with English Reuters Newswires

To provide an objective basis for comparison of our results with others, particularly [Lewis, 1992a, Lewis, 1992b], we made an extensive number of runs using English Reuters[1] data. These are 21,450 news stories from 1987. All stories beyond April 7th were used as independent test cases, and the remaining data were the training cases. The data consists of 14,704 training cases and 6,746 test cases. There are 135 topics of interest, with 93 of these topics occurring more than once in the training data. Of these newswires, there are 7133 stories with "empty" topic assignments. We chose to ignore these stories, since we can neither learn from them or test on them. As a result, the raw data that we worked with had 10,645 training cases and 3,672 test cases. We derived our own dictionaries and attributes from the raw document training data and applied rule induction machine learning methods (Swap-1). For each experiment for a given topic, a random subset, corresponding to 33% of the training data, was reserved for error estimation. Each of the recursively pruned rule sets was evaluated on these randomly selected cases to help select the best rule set. Estimates on these cases were generally within 2% of the performance of the selected rule sets on the 3,672 independent test cases from after April 7th.

wheat & farm ⟶ wheat
wheat & commodity ⟶ wheat
bushels & export ⟶ wheat
wheat & agriculture ⟶ wheat
wheat & tonnes ⟶ wheat
wheat & winter & ¬soft ⟶ wheat

	Test Cases	
	wheat	not wheat
wheat	73	8
not wheat	14	3577

Table 3: Induced Rule Set and Performance on Test Data for Reuters' English "wheat" Category

Dictionaries were created two different ways. First was the approach using the local dictionary procedure, where the 150 most frequent words for the given topic were generated. A brief universal list of stopwords was maintained, and these words were removed from the most frequent 150 words. The second approach was to create a universal dictionary by counting all words in all documents in the training set, except for stop words. Depending on the topic, a variable number of features were derived by an entropy-based feature selection method. From a universal dictionary of approximately 10,000 features, the number of features selected for each category ranged between 30 and 200. For the text representation, we experimented with both frequency and boolean features.

Performance is measured by *recall* and *precision*. Recall is the percentage of total documents for the given topic that are correctly classified. Precision is the percentage of predicted documents for the given topic that are correctly classified. Because the document topics are not mutually exclusive, document classification problems are usually analyzed as a series of dichotomous classification problems, i.e the given topic vs. not that topic. For example, Table 3 illustrates the rule set that was induced for the *wheat* category for a local dictionary with a boolean representation for the text. Also included in the figure is the performance table of this rule set on the Reuters post-April-7-1987 test data. Given the rule evaluation table as in Table 3, one can measure performance using a wide variety of metrics, based on error rates or costs. For the purpose of this study, we have chosen the *microaverage* measure, as used in [Lewis and Ringuette, 1994]. To evaluate overall performance across the entire set of topics, the results are microaveraged, i.e. the performance tables for each of the topics, such as in Table 3, are added and the overall recall and precision are computed. The point at which recall equals precision is the *breakeven* point; it can be used as a single summarizing measure for comparison of results.

The breakeven point for each of the four combinations of dictionaries and features is illustrated in Table 4. In addition, the previously reported breakeven points of 67% for decision trees [Lewis and Ringuette, 1994] and 65% for a probabilistic method [Lewis, 1992a] are listed. If all text is treated uniformly, the breakeven point for the local dictionary with frequency features is 78.9%. However, the newswire stories contain a one line headline that can provide additional clues to the topic. If the words occurring in the headline are given additional emphasis, by counting them twice, instead of a uniform count for words in either the headline or body of an article, then performance for the local dictionary with frequency features is increased by almost 2%, to a breakeven point of 80.5%.

[1] The latter was obtained by anonymous ftp from /pub/doc/reutersl on ftp.cs.umass.edu. Free distribution for research purposes has been granted by Reuters and Carnegie Group. Arrangements for access were made by David Lewis.

Learning Method	Dictionary	Text Representation	Performance Breakeven (%)
Optimized Rule Induction	Local	Frequency + Headlines	80.5
		Frequency	78.9
		Boolean	78.5
	Universal	Frequency	78.0
		Boolean	75.5
Decision Tree			67.0
Probabilistic Bayes			65.0

Table 4: Breakeven Points for Reuters' English Data

A breakeven point is a combined summary measure, but for text categorization both recall and precision may be of interest. Figure 2 illustrates the overall performance of the rule induction variations. To determine a breakeven point several learning experiments must be performed and some parameter must be varied to elicit the tradeoff of recall and precision. The appropriate technique may vary with the learning method. For rule induction, the traditional goal is to minimize the number of errors. However, this may not occur at the breakeven point. We used the standard approach of substituting costs for errors to vary the recall and precision. For a cost of one, each false negative is counted as one error[2], but for a cost of two, each false negative is counted as 2 errors. The effect of increasing the cost of false negatives is to increase the recall, at the expense of precision. In our experiments with the Reuters data, the breakeven point was achieved near a cost setting of three.

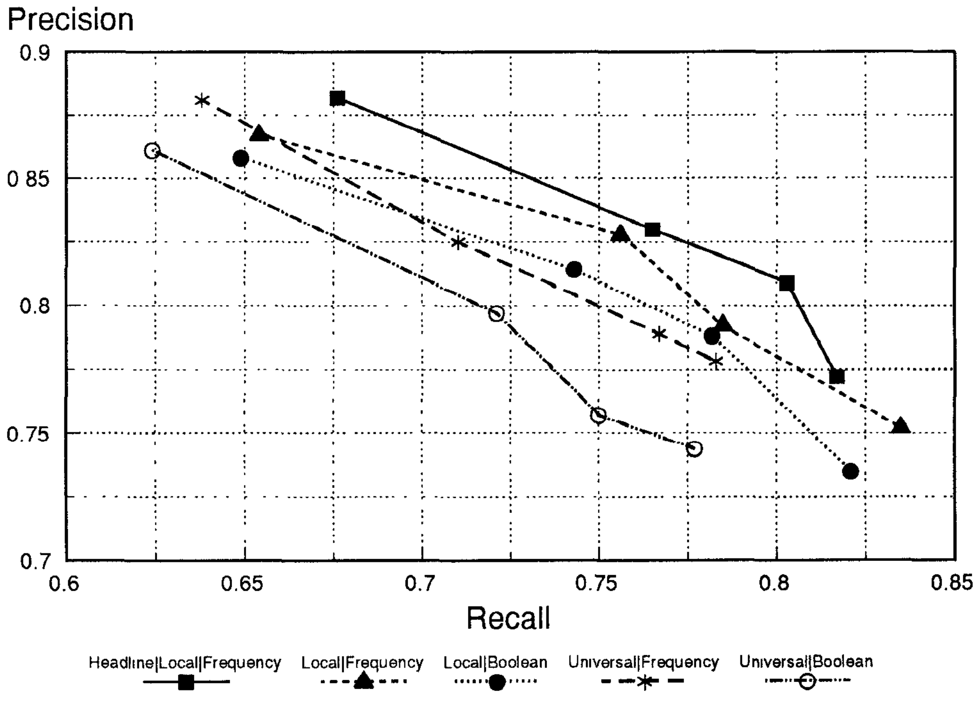

Figure 2: Recall/Precision Tradeoff for Reuters' English Data

4 Results with German Reuters Newswires

From the best results that we obtained with the English Reuters data, we observed the following salient features in our automated learning methodology:

- The usage of local dictionaries for each individual topic.
- A language independent match mechanism for determining the entries in the local dictionary.

[2] A cost of one is equivalent to the usual minimum error criterion.

- The usage of frequency of occurrence of local dictionary entries in the feature vector for representing the news stories.

- Weighting terms encountered in headlines of stories.

Given that our methodology does not rely on any language-specific technique, we hypothesized that it should perform similarly well on document collections in other languages. To provide an objective verification of this hypothesis, we have conducted detailed experiments with German Reuters[3] data. These are 8,367 news stories from 1993. All stories beyond July 1st are used as independent test cases, and the remaining data were the training cases. The data consist of 6,479 training cases and 1,888 test cases. There are 90 topics of interest.

sicherheitsrat(security council) ⟶ G131
militaerische(military) ⟶ G131
raketen(rockets) ⟶ G131
bosnien(Bosnia) ⟶ G131
nato ⟶ G131
soldaten(soldiers) ⟶ G131

	Test Cases	
	G131	not G131
G131	84	43
not G131	7	774

Table 5: Induced Rule Set and Performance on Test Data for Reuters' German "Defence" Category

Unlike the simple list of topics in the English Reuters collection, the topics for the German Reuters formed a taxonomy. There were 4 major topics, and two levels of hierarchies below the major level. At the leaf level of the hierarchy, we encountered very few stories per topic. The German Reuters topics were much broader than the ones we saw in the English Reuters. Our preliminary analysis suggested that there was much more variability in assigning topics to stories. For assigning topics from a taxonomy to a story, a complete assignment should include one or more topic from each level of the taxonomy for a story. However, analysis of our collection indicated that humans assigned fewer topics for a story than we deemed appropriate.

This poses a problem to a classification learning apparatus. Assuming that the system is attempting to learn a topic assignment model for a topic T, there will be in the training cases two sets of examples, the T set, that corresponds to all the newswires that had T as one of their topics, and the $\neg T$ set, that corresponds to all the newswires that did not have T as one of their topics. Given the disparity in topic assignment that we observed in the raw data, there will possibly be present in the $\neg T$ set, stories with topics that are either direct or indirect parents or children of T in the classification taxonomy. This will cause weakening of classification tests that otherwise would have been strong discriminators.

How can we modify our machine learning apparatus so that it can learn topic assignment models in the presence of this uncertainty? The training set that we prepare may be modified in either a simple or an extended fashion;

- simple: remove direct parents or children from the competition with their direct descendants, e.g. for topic T, remove all stories from the $\neg T$ set that have topics that are direct parents or children of T.

- extended: remove all ancestors and descendants from the competition, e.g. for topic T, remove all stories from the $\neg T$ set that have topics that are either direct or indirect parents or descendants of T.

Under the assumption that the extended option will result in categorization models that are more accurate, we ran our experiments with the extended modification option, otherwise using exactly the same apparatus that was used for obtaining the best results with English Reuters. We used a pre-stopword top-300 word local dictionary for each topic, and averaged about 125 words in the post-stopword dictionary. Table 5 illustrates an example of an induced rule set and its performance on topic G131 (Defense). The rules in this example and for other topics as well seem to be intuitively appropriate for the most part. Table 6 illustrates the breakeven performance that we obtained for some variations that have been tried so far. Note that for the lower levels in the taxonomy we tested topics that had more than 180 newswires,

[3] This collection was obtained by IBM Research from Reuters for exclusive use in this ongoing research.

and more than 100 newswires. At the top level we chose all topics since sufficient newswires were available for each of them. We can observe that the performance declines as we include topics with fewer newswires.

Topics Selected	Number of Topics	Performance Breakeven (%)	Breakeven Cost
Top level	4	73	2
Lower levels (> 180 newswires)	31	67	3
Lower levels (> 100 newswires)	43	66	3
All levels (> 180 newswires)	35	70	2-3
All levels (> 100 newswires)	47	69	2-3

Table 6: Breakeven Points for Reuters' German Data

Our current hypothesis for the slightly lower performance as compared to the English Reuters is twofold. First, the sample size that we are working with for the German collection is far smaller than the English collection. Secondly, the uncertainties introduced due to incomplete topic assignments (by the Reuters staff) from a taxonomic viewpoint have not been fully resolved in our existing methodology and will need continued investigation.

5 Discussion

Our methodology for automatic generation of text categorization models seems to perform consistently across document collections in these two languages. It is our hypothesis that our "match" mechanism for creating dictionary tokens and representing text, and the ability of the rule induction process to learn category distinguishing pattern sets using these tokens, are the two principle components that provide this pattern-directed language-independent classification power.

When compared to previous results on the English Reuters data, our new results appear significantly better. Figure 2 suggests that the use of local dictionaries and frequency information were effective and improved the results of our rule induction methods. By far the greatest improvement came from the learning method (Swap-1). While previous experience has shown that the optimization techniques of this rule induction method can often substantially improve results over competitive methods, such as decision trees, text classification has a number of characteristics that make optimized rule induction particularly suitable. The optimization techniques that are employed are quite strong in finding feature dependencies. In terms of text classification this means that given single word dictionaries it can find the key word combination co-occurrences that distinguish between topics.

A source of uncertainty with the German newswires is caused by the hierarchic classification system in use. For the top level classes, there is no problem. However, there is no uniform pattern of human code assignment below the top level. That is, some stories have only a top level code assigned. How is this to be compared to a story with a leaf node code assigned which lies in the same hierarchy? Is the assignment of only a top level code to be considered equivalent to a leaf node code of "other"?. If not, how should such stories be used when computing classification results on the test data set? These issues require continued investigation.

German as a language has two salient major differences from English (although they are both in the same language family and therefore have many resemblances). The first is that German is a more heavily inflected language and one might suppose that it is more important to normalize word forms to stems. Our results do not seem to bear this out, however. It appears that the frequent non-common words which we use as features tend to occur with the same inflection in a topic, for the most part. The second major difference is that German tends to use word formations involving several stems to form a single word where English would use a noun phrase. One might suppose that these would need to be decomposed. Again, our results do not support this. Relatively few of the long compounds occur as features, and those that do might be considered to be specialized idioms which should not be decomposed. For example, in one experiment using the top 150 non-common word stems as features for the category "politics", the only compounds were "aussenminister" = "exterior minister", "menschenrecht" = "human rights", "ministerpraesident" = "minister president", and "mitgliedstaaten" = "partner states", all arguably idioms in this context.

From these experiments, it appears that optimized rule induction is more than competitive with other machine learning techniques [Masand *et al.*, 1992, Lewis and Ringuette, 1994, Lewis, 1992a] for document classification, and very close behind human-engineered systems [Hayes and Weinstein, 1991]. Such conclusions can only be supported by rigorous and exacting comparisons. The 1987 English Reuters stories have recently been widely circulated and should prove to be an important benchmark for objective comparisons. The 1993 German Reuters collection will hopefully provide an additional benchmark as we and others continue to extend our results.

References

[Apté *et al.*, 1993] C. Apté, F. Damerau, and S. Weiss. Automated Learning of Decison Rules for Text Categorization. Technical Report RC 18879, IBM T.J. Watson Research Center, 1993. To appear in ACM Transactions on Office Information Systems.

[Breiman *et al.*, 1984] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees.* Wadsworth, Monterrey, Ca., 1984.

[Hayes and Weinstein, 1991] P. Hayes and S. Weinstein. Adding Value to Financial News by Computer. In *Proceedings of the First International Conference on Artificial Intelligence Applications on Wall Street*, pages 2–8, 1991.

[Hayes *et al.*, 1990] P.J. Hayes, P.M. Andersen, I.B. Nirenburg, and L.M. Schmandt. TCS: A Shell for Content-Based Text Categorization. In *Proceedings of the Sixth IEEE CAIA*, pages 320–326, 1990.

[Lewis and Ringuette, 1994] D. Lewis and M. Ringuette. A comparison of two learning algorithms for text categorization. In *Symposium on Document Analysis and Information Retrieval*, Las Vegas, NV, April 1994. ISRI; Univ. of Nevada, Las Vegas. To appear.

[Lewis, 1992a] D. Lewis. An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task. In *Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 37–50, June 1992. Edited by Nicholas Belkin, Peter Ingwersen, and Annelise Mark Pejtersen.

[Lewis, 1992b] D. Lewis. Feature Selection and Feature Extraction for Text Categorization. In *Procceedings of the Speech and Natural language Workshop*, pages 212–217, February 1992. Sponsored by the Defense Advanced Research Projects Agency.

[Masand *et al.*, 1992] B. Masand, G. Linoff, and D. Waltz. Classifying News Stories using Memory Based Reasoning. In *Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 59–65, June 1992. Edited by Nicholas Belkin, Peter Ingwersen, and Annelise Mark Pejtersen.

[Quinlan, 1993] J.R. Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann, 1993.

[Weiss and Indurkhya, 1993] S. Weiss and N. Indurkhya. Optimized Rule Induction. *IEEE EXPERT*, 8(6):61–69, December 1993.

[Weiss and Kulikowski, 1991] S.M. Weiss and C.A. Kulikowski. *Computer Systems That Learn.* Morgan Kaufmann, 1991.