Feature Selection and Feature Extraction for Text Categorization

David D. Lewis

Center for Information and Language Studies
University of Chicago
Chicago, IL 60637

ABSTRACT

The effect of selecting varying numbers and kinds of features for use in predicting category membership was investigated on the Reuters and MUC-3 text categorization data sets. Good categorization performance was achieved using a statistical classifier and a proportional assignment strategy. The optimal feature set size for word-based indexing was found to be surprisingly low (10 to 15 features) despite the large training sets. The extraction of new text features by syntactic analysis and feature clustering was investigated on the Reuters data set. Syntactic indexing phrases, clusters of these phrases, and clusters of words were all found to provide less effective representations than individual words.

1. Introduction

Text categorization—the automated assigning of natural language texts to predefined categories based on their content—is a task of increasing importance. Its applications include indexing texts to support document retrieval [1], extracting data from texts [2], and aiding humans in these tasks.

The indexing language used to represent texts influences how easily and effectively a text categorization system can be built, whether the system is built by human engineering, statistical training, or a combination of the two. The simplest indexing languages are formed by treating each word as a feature. However, words have properties, such as synonymy and polysemy, that make them a less than ideal indexing language. These have motivated attempts to use more complex feature extraction methods in text retrieval and text categorization tasks.

If a syntactic parse of text is available, then features can be defined by the presence of two or more words in particular syntactic relationships. We call such a feature a *syntactic indexing phrase*. Another strategy is to use cluster analysis or other statistical methods to detect closely related features. Groups of such features can then, for instance, be replaced by a single feature corresponding to their logical or numeric sum. This strategy is referred to as *term clustering*.

Syntactic phrase indexing and term clustering have opposite effects on the properties of a text representation, which led us to investigate combining the two techniques [3]. However, the small size of standard text retrieval test collections, and the variety of approaches available for query interpretation, made it difficult to study purely representational issues in text retrieval experiments. In this paper we examine indexing language properties using two text categorization data sets. We obtain much clearer results, as well as producing a new text categorization method capable of handing multiple, overlapping categories.

2. Data Sets and Tasks

Our first data set was a set of 21,450 Reuters newswire stories from the year 1987 [4]. These stories have been manually indexed using 135 financial topic categories, to support document routing and retrieval. Particular care was taken in assigning categories [1]. All stories dated April 7, 1987 and earlier went into a set of 14,704 training documents, and all stories from April 8, 1987 or later went into a test set of 6,746 documents.

The second data set consisted of 1,500 documents from the U.S. Foreign Broadcast Information Service (FBIS) that had previously been used in the MUC-3 evaluation of natural language processing systems [2]. The documents are mostly translations from Spanish to English, and include newspaper stories, transcripts of broadcasts, communiques, and other material.

The MUC-3 task required extracting simulated database records ("templates") describing terrorist incidents from these texts. Eight of the template slots had a limited number of possible fillers, so a simplification of the MUC-3 task is to view filling these slots as text categorization. There were 88 combinations of these 8 slots and legal fillers for the slots, and each was treated as a binary category. Other text categorization tasks can be defined for the MUC-3 data (see Riloff and Lehnert in this volume).

We used for our test set the 200 official MUC-3 test documents, plus the first 100 training documents (DEV-MUC3-0001 through DEV-MUC3-0100). Templates for these 300 documents were encoded by the MUC-3 orga-

nizers. We used the other 1,200 MUC-3 training documents (encoded by 16 different MUC-3 sites) as our categorization training documents. Category assignments should be quite consistent on our test set, but less so on our training set.

3. Categorization Method

The statistical model used in our experiments was proposed by Fuhr [5] for probabilistic text retrieval, but the adaptation to text categorization is straightforward. Figure 1 shows the formula used. The model allows the possibility that the values of the binary features for a document is not known with certainty, though that aspect of the model was not used in our experiments.

3.1. Binary Categorization

In order to compare text categorization output with an existing manual categorization we must replace probability estimates with explicit binary category assignments. Previous work on statistical text categorization has often ignored this step, or has not dealt with the case where documents can have zero, one, or multiple correct categories.

Given accurate estimates of $P(C_j = 1|D_m)$, decision theory tells us that the optimal strategy, assuming all errors have equal cost, is to set a single threshold p and assign C_j to a document exactly when $P(C_j = 1|D_m) >= p$ [6]. However, as is common in probabilistic models for text classification tasks, the formula in Figure 1 makes assumptions about the independence of probabilities which do not hold for textual data. The result is that the estimates of $P(C_j = 1|D_m)$ can be quite inaccurate, as well as inconsistent across categories and documents.

We investigated several strategies for dealing with this problem and settled on *proportional assignment* [4]. Each category is assigned to its top scoring documents on the test set in a designated multiple of the percentage of documents it was assigned to on the training corpus. Proportional assignment is not very satisfactory from a theoretical standpoint, since the probabilistic model is supposed to already take into account the prior probability of a category. In tests the method was found to perform well as a standard decision tree induction method, however, so it is at least a plausible strategy. We are continuing to investigate other approaches.

3.2. Feature Selection

A primary concern of ours was to examine the effect of feature set size on text categorization effectiveness. All potential features were ranked for each category by expected mutual information [7] between assignment of that feature and assignment of that category. The top k features for each category were chosen as its feature set, and different values of k were investigated.

- WORDS-DF2: Starts with all words tokenized by *parts*. Capitalization and syntactic class ignored. Stopwords discarded based on syntactic tags. Tokens consisting solely of digits and punctuation removed. Words occurring in fewer than 2 training documents removed. Total terms: 22,791.
- WC-MUTINFO-135: Starts with WORDS-DF2, and discards words occurring in fewer than 5 or more than 1029 (7%) training documents. RNN clustering used 135 metafeatures with value equal to mutual information between presence of the word and presence of a manual indexing category. Result is 1,442 clusters and 8,506 singlets, for a total of 9,948 terms.
- PHRASE-DF2: Starts with all simple noun phrases bracketed by parts. Stopwords removed from phrases based on tags. Single word phrases discarded. Numbers replaced with the token NUMBER. Phrases occurring in fewer than 2 training documents removed. Total terms: 32,521.
- PC-W-GIVEN-C-44: Starts with PHRASE-DF2. Phrases occurring in fewer than 5 training documents removed. RNN clustering uses 44 metafeatures with value equal to our estimate of $P(W = 1|C = 1)$ for phrase W and category C. Result is 1,883 clusters and 1,852 singlets, for a total of 3,735 terms.

Figure 2: Summary of indexing languages used with the Reuters data set.

4. Indexing Languages

We investigated phrasal and term clustering methods only on the Reuters collection, since the smaller amount of text made the MUC-3 corpus less appropriate for clustering experiments. For the MUC-3 data set a single indexing language consisting of 8,876 binary features was tested, corresponding to all words occurring in 2 or more training documents. The original MUC-3 text was all capitalized. Stop words were not removed.

For the Reuters data we adopted a conservative approach to syntactic phrase indexing. The phrasal indexing language consisted only of simple noun phrases, i.e. head nouns and their immediate premodifiers. Phrases were formed using *parts*, a stochastic syntactic class tagger and simple noun phrase bracketing program [8]. Words

We estimate $P(C_j = 1|D_m)$ by:

$$P(C_j = 1) \times \prod_i \left(\frac{P(W_i = 1|C_j = 1) \times P(W_i = 1|D_m)}{P(W_i = 1)} + \frac{P(W_i = 0|C_j = 1) \times P(W_i = 0|D_m)}{P(W_i = 0)} \right)$$

Explanation:

- $P(C_j = 1|D_m)$ is the probability that category C_j is assigned to document D_m. Estimating this probability is the goal of the categorization procedure. The index j ranges over categories to be assigned.
- $P(C_j = 1)$ is the prior probability that category C_j is assigned to a document, in the absence of any information about the contents of the particular document.
- $P(W_i = 1)$ is the prior probability that feature W_i is present in a randomly selected document. $P(W_i = 0) = 1 - P(W_i = 1)$. The index i ranges over the set of predictor features for category C_j.
- $P(W_i = 1|C_j = 1)$ is the probability that feature W_i is assigned to a document given that we know category C_j is assigned to that document. $P(W_i = 0|C_j = 1)$ is $1 - P(W_i = 1|C_j = 1)$.
- $P(W_i = 1|D_m)$ is the probability that feature W_i is assigned to document D_m.
- All probabilities were estimated from the training corpus using the "add one" adjustment (the Jeffreys prior).

Figure 1: Probabilistic model used for text categorization.

that were tagged as function words were removed from phrases, and all items tagged as numbers were replaced with the token NUMBER. We also used the *parts* segmentation to define the set of words indexed on.

Reciprocal nearest neighbor clustering was used for clustering features. An RNN cluster consists of two items, each of which is the nearest neighbor of the other according to the similarity metric in use. Therefore, not all items are clustered. If this stringent clustering strategy does not bring together closely related features, it is unlikely that any clustering method using the same metafeatures would do so.

Clustering features requires defining a set of metafeatures on which the similarity of the features will be judged. We experimented with forming clusters from words under three metafeature definitions, and from phrases under eight metafeature definitions [4]. Metafeatures were based on presence or absence of features in documents, or on the strength of association of features with categories of documents. In all cases, similarity between metafeature vectors was measured using the cosine correlation. The sets of clusters formed were examined by the author, and categorization experiments were run with the three sets of word clusters and with the two sets of phrase clusters that appeared best. Figure 2 summarizes the properties of the most effective version of each representation type used in the experiments on the Reuters data.

5. Evaluation

The effectiveness measures used were *recall* (number of categories correctly assigned divided by the total number of categories that should be assigned) and *precision* (number of categories correctly assigned divided by total number of categories assigned).

For a set of k categories and d documents a total of $n = kd$ categorization decisions are made. We used *microaveraging*, which considers all kd decisions as a single group, to compute average effectiveness [9]. The proportionality parameter in our categorization method was varied to show the possible tradeoffs between recall and precision. As a single summary figure for recall precision curves we took the *breakeven* point, i.e. the highest value (interpolated) at which recall and precision are equal.

6. Results

We first looked at effectiveness of proportional assignment with word-based indexing languages. Figure 3 shows results for the best feature set sizes found: 10 features on Reuters and 15 features on MUC-3. A breakeven point of 0.65 on Reuters and 0.48 on MUC-3 is reached. For comparison, the operational AIR/X system uses both rule-based and statistical techniques to achieve a microaveraged breakeven point of approximately 0.65 in indexing a physics database [10].

The CONSTRUE rule-based text categorization system achieves a microaveraged breakeven of around 0.90 on

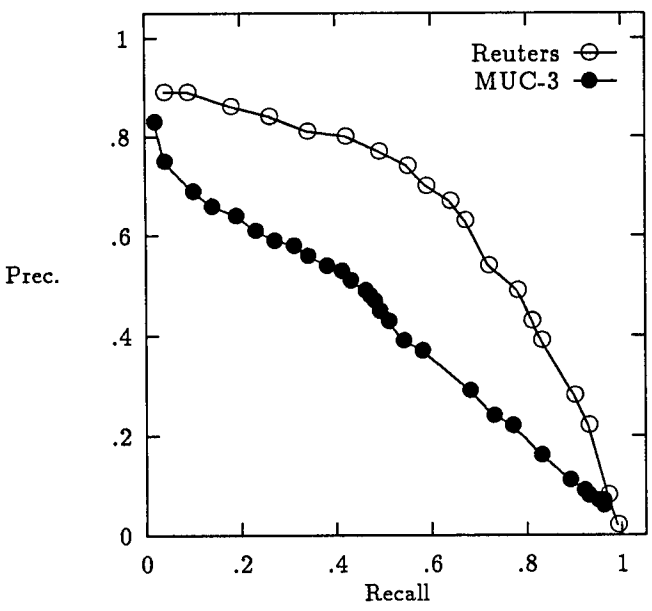

Figure 3: Microaveraged recall and precision on Reuters (w/ 10 features) and MUC-3 (w/ 15 features) test sets.

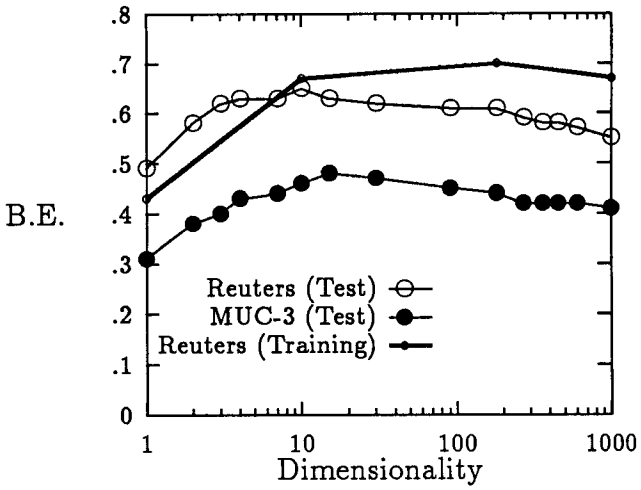

Figure 4: Microaveraged breakeven points for feature sets of words on Reuters and MUC-3 test sets, and on Reuters training set.

a different, and possibly easier, testset drawn from the Reuters data [1]. This level of performance, the result of a 9.5 person-year effort, is an admirable target for learning-based systems to shoot for.

Comparison with published results on MUC-3 are difficult, since we simplified the complex MUC-3 task. However, in earlier experiments using the official MUC-3 testset and scoring, proportional assignment achieved performance toward but within the low end of official MUC-3 scores achieved by a variety of NLP methods. This is despite being limited in most cases to 50% the score achievable by methods that attempted cross-referencing [11].

6.1. Feature Selection

Figure 4 summarizes our data on feature set size. We show the breakeven point reached for categorization runs with various size sets of words, again on both the Reuters and MUC-3 data sets. The results exhibit the classic peak associated with the "curse of dimensionality." The surprise is the small number of features found to be optimal. With 14,704 and 1,300 training examples, peaks of 10 and 15 features respectively are smaller than one would expect based on sample size considerations.

Overfitting, i.e. training a model on accidental as well as systematic relationships between feature values and category membership, was one possible villain [6]. We checked for overfitting directly by testing the induced classifiers on the training set. The thicker line in Figure 4 shows the effectiveness of the Reuters classifiers when tested on the 14,704 stories used to train them. Surprisingly, effectiveness reaches a peak not much higher than that achieved on the unseen test set, and even drops off when a very large feature set is used. Apparently our probabilistic model is sufficiently constrained that, while overfitting occurs, its effects are limited.[1]

Another possible explanation for the decrease in effectiveness with increasing feature set size is that the assumptions of the probabilistic model are increasingly violated. Fuhr's model assumes that the probability of observing a word in a document is independent of the probability of observing any other word in the document, both for documents in general and for documents known to belong to particular categories. The number of opportunities for groups of dependent features to be selected as predictor features for the same category increases as the feature set size grows.

Finally, since features with a higher value on expected mutual information are selected first, we intuitively expect features with lower ratings, and thus appearing only in the larger feature sets, to simply be worse features. This intuition is curiously hard to justify. Any feature has some set of conditional and unconditional probabilities and, if the assumptions of the statistical model hold,

[1] We have not yet made this test on the MUC-3 data set.

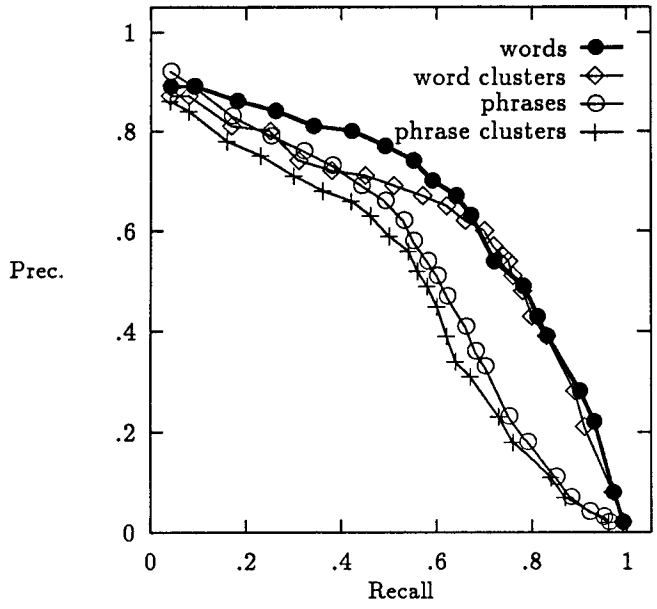

Figure 5: Microaveraged recall and precision on Reuters test set for WORDS-DF2 words (10 features), WC-MUTINFO-135 word clusters (10 features), PHRASE-DF2 phrases (180 features), and PC-W-GIVEN-C-44 phrase clusters (90 features).

will be used in an appropriate fashion. It may be that the inevitable errors in estimating probabilities from a sample are more harmful when a feature is less strongly associated with a category.

6.2. Feature Extraction

The best results we obtained for each of the four basic representations on the Reuters test set are shown in Figure 5. Individual terms in a phrasal representation have, on the average, a lower frequency of appearance than terms in a word-based representation. So, not surprisingly, effectiveness of a phrasal representation peaks at a much higher feature set size (around 180 features) than that of a word-based representation (see Figure 6). More phrases are needed simply to make any distinctions among documents. Maximum effectiveness of the phrasal representation is also substantially lower than that of the word-based representation. Low frequency and high degree of synonymy outweigh the advantages phrases have in lower ambiguity.

Disappointingly, as shown in Figure 5, term clustering did not significantly improve the quality of either a word-based or phrasal representation. Figure 7 shows some representative PC-W-GIVEN-C-44 phrase clusters.

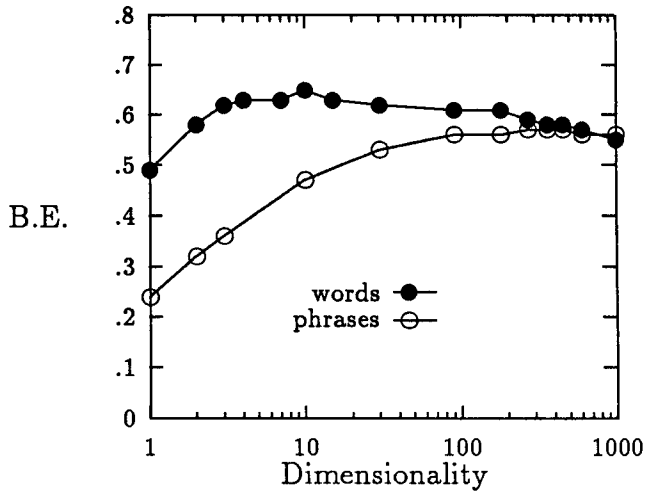

Figure 6: Microaveraged breakeven point for various sized feature sets of words and phrases on Reuters test set.

(The various abbreviations and other oddities in the phrases were present in the original text.) Many of the relationships captured in the clusters appear to be accidental rather than the systematic semantic relationships hoped for.

Why did phrase clustering fail? In earlier work on the CACM collection [3], we identified lack of training data as a primary impediment to high quality cluster formation. The Reuters corpus provided approximately 1.5 million phrase occurrences, a factor of 25 more than CACM. Still, it remains the case that the amount of data was insufficient to measure the distributional properties

's investors service inc, < amo >
NUMBER accounts, state regulators
NUMBER elections, NUMBER engines
federal reserve chairman paul volcker,
private consumption
additional NUMBER dlrs, america >
canadian bonds, cme board
denmark NUMBER, equivalent price
fund government-approved equity investments,
fuji bank ltd
its share price, new venture
new policy, representative offices
same-store sales, santa rosa

Figure 7: Some representative PC-W-GIVEN-C-44 clusters.

of many phrases encountered.

The definition of metafeatures is a key issue to reconsider. Our original reasoning was that, since phrases have low frequency, we should use metafeatures corresponding to bodies of text large enough that we could expect cooccurrences of phrases within them. The poor quality of the clusters formed suggests that this approach is not effective. The use of such coarse-grained metafeatures simply gives many opportunities for accidental cooccurrences to arise, without providing a sufficient constraint on the relationship between phrases (or words). The fact that clusters captured few high quality semantic relationships, even when an extremely conservative clustering method was used, suggests that using other clustering methods with the same metafeature definitions is not likely to be effective.

Finally, while phrases are less ambiguous than words, they are not all good content indicators. Even restricting phrase formation to simple noun phrases we see a substantial number of poor content indicators, and the impact of these are compounded when they are clustered with better content indicators.

7. Future Work

A great deal of research remains in developing text categorization methods. New approaches to setting appropriate category thresholds, estimating probabilities, and selecting features need to be investigated. For practical systems, combinations of knowledge-based and statistical approaches are likely to be the best strategy.

On the text representation side, we continue to believe that forming groups of syntactic indexing phrases is an effective route to better indexing languages. We believe the key will be supplementing statistical evidence of phrase similarity with evidence from thesauri and other knowledge sources, along with using metafeatures which provide tighter constraints on meaning. Clustering of words and phrases based on syntactic context is a promising approach (see Strzalkowski in this volume). Pruning out of low quality phrases is also likely to be important.

8. Summary

We have shown a statistical classifier trained on manually categorized documents to achieve quite effective performance in assigning multiple, overlapping categories to documents. We have also shown, via studying text categorization effectiveness, a variety of properties of indexing languages that are difficult or impossible to measure directly in text retrieval experiments, such as effects of feature set size and performance of phrasal representations in isolation from word-based representations.

Like text categorization, text retrieval is a text classification task. The results shown here for text categorization, in particular the ineffectiveness of term clustering with coarse-grained metafeatures, are likely to hold for text retrieval as well, though further experimentation is necessary.

9. Acknowledgments

Thanks to Bruce Croft, Paul Utgoff, Abe Bookstein and Marc Ringuette for helpful discussions. This research was supported at U Mass Amherst by grant AFOSR-90-0110, and at the U Chicago by Ameritech. Many thanks to Phil Hayes, Carnegie Group, and Reuters for making available the Reuters data, and to Ken Church and AT&T for making available *parts*.

References

1. Hayes, P. and Weinstein, S. CONSTRUE/TIS: a system for content-based indexing of a database of news stories. In *IAAI-90*, 1990.
2. Sundheim, B., ed. *Proceedings of the Third Message Understanding Evaluation and Conference*, Morgan Kaufmann, Los Altos, CA, May 1991.
3. Lewis, D. and Croft, W. Term clustering of syntactic phrases. In *ACM SIGIR-90*, pp. 385–404, 1990.
4. Lewis, D. *Representation and Learning in Information Retrieval.* PhD thesis, Computer Science Dept.; Univ. of Mass.; Amherst, MA, 1992. Technical Report 91–93.
5. Fuhr, N. Models for retrieval with probabilistic indexing. *Information Processing and Management*, 25(1):55–72, 1989.
6. Duda, R. and Hart, P. *Pattern Classification and Scene Analysis.* Wiley-Interscience, New York, 1973.
7. Hamming, R. *Coding and Information Theory.* Prentice-Hall, Englewood Cliffs, NJ, 1980.
8. Church, K. A stochastic parts program and noun phrase parser for unrestricted text. In *Second Conference on Applied NLP*, pp. 136–143, 1988.
9. Lewis, D. Evaluating text categorization. In *Speech and Natural Language Workshop*, pp. 312–318, Feb. 1991.
10. Fuhr, N., et al. AIR/X—a rule-based multistage indexing system for large subject fields. In *RIAO 91*, pp. 606–623, 1991.
11. Lewis, D. Data extraction as text categorization: An experiment with the MUC-3 corpus. In *Proceedings MUC-3*, May 1991.