Text Categorization: An Experiment Using Phrases

Madhusudhan Kongovi, Juan Carlos Guzman, and Venu Dasigi

Southern Polytechnic State University
1100 S. Marietta Parkway
Marietta, GA 30060
mkongovi@lucent.com
jguzman@spsu.edu
vdasigi@spsu.edu

Abstract. Typical text classifiers learn from example and training documents that have been manually categorized. In this research, our experiment dealt with the classification of news wire articles using category profiles. We built these profiles by selecting feature words and phrases from the training documents. For our experiments we decided on using the text corpus Reuters-21578. We used precision and recall to measure the effectiveness of our classifier. Though our experiments with words yielded good results, we found instances where the phrase-based approach produced more effectiveness. This could be due to the fact that when a word along with its adjoining word – a phrase – is considered towards building a category profile, it could be a good discriminator. This tight packaging of word pairs could bring in some semantic value. The packing of word pairs also filters out words occurring frequently in isolation that do not bear much weight towards characterizing that category.

1. Introduction

Categorization is a problem that cognitive psychologists have dealt with for many years [1]. There are two general and basic principles for creating categories: *cognitive economy* and *perceived world structure* [12]. The principle of cognitive economy means that the function of categories is to provide maximum information with the least cognitive effort. The principle of perceived world structure means that the perceived world is not an unstructured set of arbitrary or unpredictable attributes. The attributes that an individual will perceive, and thus use for categorization, are determined by the needs of the individual. These needs change over time and with the physical and social environment. In other words, a system for automatic text categorization should in some way "know" both the type of text and the type of user. The maximum information with least cognitive effort is achieved if categories map the perceived world structure as closely as possible (*ibid*). Coding by category is fundamental to mental life because it greatly reduces the demands on perceptual processes, storage space, and reasoning processes, all of which are known to be limited [13]. Psychologists agree that *similarity* plays a central role in placing

F. Crestani, M. Girolami, and C.J. van Rijsbergen (Eds.): ECIR 2002, LNCS 2291, pp. 213-228, 2002.

different items into a single category. Furthermore, people want to maximize within-category similarity while minimizing between-category similarity (*ibid*).

In the process of categorization of electronic documents, categories are typically used as a means of organizing and getting an overview of the information in a collection of several documents. *Folders* in electronic mail (e-mail) and *topics* in Usenet News are a couple of concrete examples of categories in computer-mediated communication.Text categorization is, in this paper, defined as an information retrieval task in which one category label is assigned to a document [11]. Techniques for automatically deriving representations of categories ("category profile extraction") and performing classification have been developed within the area of text categorization [9], a discipline at the crossroads between information retrieval and machine learning. Alternatively, a document can be compared to previously classified documents and placed in the category of the closest such documents [6], avoiding the need for category profiles. All these categorization approaches perform *categorization by content* [14], since information for categorizing a document is extracted from the document itself.

Our work primarily focuses on building a category profile of words and phrases as a vocabulary for a category and using that to perform a match, to categorize.

Terms & Definitions

Feature Words – Representative words that describe a given information category. Words that occur more regularly and more frequently than others in documents of a category are good candidates for topic words. These words are different from *stop words*, which are defined below.

Stop words are words with little or no indexing value and would comprise conjunctions, prepositions, adverbs, articles, some verbs, pronouns and some proper names. Although we have tried to come up with a standard, there is some unavoidable subjectivity.

Phrases – two adjoining words in the text with zero word distance, eliminating all the stop words in between.

The retrieval activity divides the collection into four parts, consisting of relevant retrieved items (a), relevant not retrieved (d), non-relevant retrieved (b) and non-relevant not retrieved (c).

True Positive (a_i) **TP** - This represents the total number of relevant documents retrieved for a particular category (i).

False Positive (b_i) **FP** - This represents the total number of non-relevant documents retrieved for a particular category (i).

True Negative (c_i) **TN** - This represents the number of non-relevant documents not retrieved for a particular category (i).

False Negative (d_i) **FN** - This represents the total number of relevant documents not retrieved for a particular category (i).

We can define precision for that category (i) as follows:
$Precision_i = a_i /(a_i + b_i)$
$Recall_i = a_i /(a_i + d_i)$

The overall performance measures for a collection having C categories can be defined [Dasigi et al, 2001] as follows:

Macro precision = $\sum_i Precision_i / C$

Macro recall = $\sum_i Recall_i / C$

Micro precision = $\sum_i a_i / \sum_i (a_i + b_i)$

Micro recall = $\sum_i a_i / \sum_i (a_i + d_i)$

Typicality is a measure of how typical or how representative a document is of a particular category. Suppose we identify 15 phrases or words that characterize a particular category. And we decide that a document needs to have all the 15 phrases or words to be 100% representative of that category and if a document **A** contains 12 phrases or words, its typicality is 12/15 or 0.8 or it is 80% typical of that document. And the **semantic distance** of the two documents, in that category, is measured by the difference of their typicalities in that category.

2. Previous Related Research

The basic approach of Verma and Dasigi [3] was to first come up with a pattern of words (called category profiles) defining a given information category, and then apply classification algorithms that make use of these profiles to classify test documents. David Lewis's primary research area is information retrieval, including the categorization, retrieval, routing, filtering, clustering, linking, etc. of text. His [7, 8] research focuses on the application of statistical and machine learning techniques in information retrieval (IR). Automatic text categorization is an important research area in Katholieke Universiteit Leuven, Belgium. Marie-Francine Moens and Jos Dumortier have worked on applying text categorization techniques in new areas of text routing and filtering. Their research discusses the categorization of magazine articles with broad subject descriptors. They especially focus upon the following aspects of text classification: effective selection of feature words and proper names that reflect the main topics of the text, and training of text classifiers. Fabrizio Sebastiani of Instituto di Elaborazione dell'Informazione [5], Pisa, Italy, investigates a novel technique for automatic categorization, which is dubbed *categorization by context*, since it exploits the context surrounding a link in an HTML document to

extract useful information for categorizing the document referred by the link. Neural networks have proven very effective for tasks involving pattern recognition since text classification, in essence, is a pattern recognition problem. Some researches [2] have used neural networks in conjunction with the latent semantic indexing model [4] to classify text data.

3. Text Corpus

As our work was motivated by a previous research [3] in text classification using the corpus Reuters-21578, we decided to use the same source for our experiment so that we could compare results and determine if our work added any value to the existing body of research literature. For our experiments we decided on working with five categories – GRAIN, CRUDE, METAL, MONEY-FX and SHIP. As we set out with the idea that our experiments – categorization task – should result in categorizing test documents in just one category, we decided on using stories – both for training and testing – that had less chance of falling into more than one category.

4. Empirical Work: An Attempt to Automate Classification Based on Phrases

Our empirical research consisted of developing an automatic text classifier. Considering the scope of the experiment, the following assumptions were made:

i. The training and test documents would only come from the corpus Reuters-21578. No other source of data would be used. The documents would distinctly fall under just one category.

ii. We would use a much larger set to train the classifier as compared to testing the classifier.

iii. We define a phrase as two adjoining words in the text with zero word distance, eliminating all the stop words.

iv. Stop words would comprise conjunctions, prepositions, adverbs, articles, some verbs, pronouns and some proper names. Although we have tried to come up with a standard, there is some unavoidable subjectivity.

v. Lastly, as this is an experimental prototype, we did not spend much time on writing efficient programs using sophisticated software tools and utilities.

Programs

Since our work environment was UNIX, we used a lot of tools and utilities available in the UNIX world. The core programs were written in the C language. The lexical analyzer was written using the tool LEX. The computation of results and manipulation were done using the AWK and SHELL programming languages. The final analysis and plotting of graphs were done using EXCEL. As the Reuters-21578 data is in SGML format, a program had to be written to filter out the SGML tags and extract the text along with the topic names for the training document set. We had decided on using five categories for the experiment – GRAIN, CRUDE, MONEY-FX, METAL and SHIP. After extraction, the training documents were put in separate directories. Table 1 shows the number of test and training documents for each category.

Table 1. Subject categories and number of test and training documents

Category	Category Description	# Test Docs	# Train Docs
Grain	All topics that include grain	58	324
Crude	Oil Prices, Production & Distribution	226	2619
Money-Fx	Money Market & Exchange	47	501
Ship	Shipping	58	154
Metal	All topics that include Metal	60	64

We came up with a list of stop words – words that do not add value towards categorization – consisting of adverbs, verbs, conjunctions, prepositions, adjectives, pronouns and some proper names. The primary task involved extracting the phrases from the training documents. The lex program generated the tokens (phrases) based on the rules – stop words, sentence beginnings and categories – we had provided. The phrases across all the training documents, specific to the categories, were collected. This resulted in five buckets – categories – of unsorted phrases. The next step was to compute the frequencies of the phrases. The phrase lists were further processed based on the relative frequencies of the phrases and probabilities. The list was ordered highest to the lowest frequency. The last step involved maintaining a column of cumulative probabilities in the ordered list. Thus the final ordered phrases list per category contained five columns comprising **phrase**, **frequency**, **relative frequency**, **probability** and **cumulative probability**. These lists gleaned from the training documents formed the phrases-vocabulary for the classifier. The classifier would look up these lists to find phrase matches in the test documents in the process of classifying. The category with the highest match is what the test document would belong to. A document would not be categorized (**No Cat**) if it got equal score in more than one category or did not get the minimum score in any.

Experiments

Since there was no clear picture of how many phrases would fairly represent a category, we decided to conduct a series of experiments using the programs mentioned in the previous section.

Absolute Cut-off

We used an absolute number of phrases per category from their respective phrase list to start with. Of the list of the phrases – by frequency – we picked the top 15 from each category and used that as the look up table. Figure 1 shows the precision and recall values for this experiment. Recall is poor as quite a big number of documents did not get categorized, resulting in no category (**No CAT**).

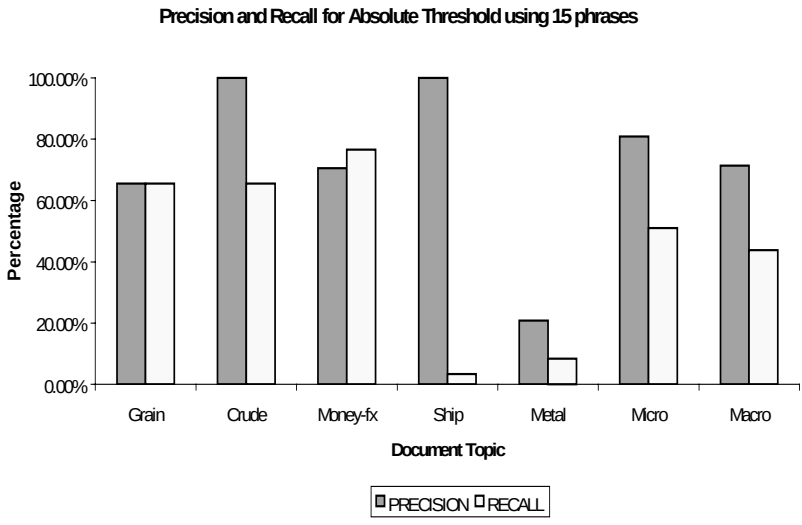

Fig. 1. Precision and recall values for an absolute threshold of 15 phrases

Relative Cut-off

For each category, with the phrases sorted in decreasing order of probability, the cumulative probability for the top ranked phrases down to each ranked position is kept track of. When this cumulative probability is reached at a certain chosen threshold point, we cut off the selection of phrases for that category. For the first experiment we used top ranked phrases until the cumulative probability (**CP**) totaled up to **0.08** as the cut-off. We had varied number of phrases across different categories. Table 2 shows the results of this run.

Table 2. Results of the 0.08CP run

Category	# Test Docs	TP	FP	No CAT	FN
Grain	58	49	16	5	4
Crude	226	163	0	31	32
Money-Fx	47	45	13	1	1
Ship	58	36	13	17	5
Metal	60	20	17	24	16

For the next experiment we raised the cumulative probability threshold to **0.1** as the cut-off. The number of phrases naturally went up as shown in Table 3.

Table 3. Results of the 0.1CP run

Category	# Test Docs	TP	FP	No CAT	FN
Grain	58	49	12	4	5
Crude	226	201	3	15	10
Money-Fx	47	46	7	1	0
Ship	58	37	11	14	7
Metal	60	22	7	24	14

For the next experiment we raised the cut-off threshold to **0.13** cumulative probability as shown in Table 4.

Table 4. Results of the 0.13CP run

Category	# Test Docs	TP	FP	No CAT	FN
Grain	58	50	19	3	5
Crude	226	180	0	20	26
Money-Fx	47	44	19	1	2
Ship	58	39	8	12	7
Metal	60	37	10	11	12

The results of the experiment using a **0.18** cumulative probability threshold as the cut-off are shown in Table 5.

Table 5. Results of the 0.18CP run

Category	# Test Docs	TP	FP	No CAT	FN
Grain	58	53	32	1	4
Crude	226	157	0	18	51
Money-Fx	47	46	27	0	1
Ship	58	39	7	10	9
Metal	60	33	16	13	14

For the final experiment we had raised the cut-off threshold to **0.25** cumulative probability as shown in Table 6.

Table 6. Results of the 0.25CP run

Category	# Test Docs	TP	FP	No CAT	FN
Grain	58	53	29	2	3
Crude	226	182	0	15	29
Money-Fx	47	47	23	0	0
Ship	58	38	6	9	11
Metal	60	30	7	11	19

Table 7. Number of phrases that qualified for different runs

Category	# of phrases	# of distinct phrases	0.08CP	0.1CP	0.18 CP	0.25 CP
Grain	34544	25514	142	229	787	1661
Crude	148416	69328	14	24	47	143
Money-Fx	58708	35349	80	133	522	1239
Ship	15125	11979	179	250	696	1225
Metal	7152	5995	104	153	414	664

Results

The analysis of the various experiments conducted shows a pattern. Table 7 shows the number of phrases that were used in different runs. The absolute cut-off threshold has the least acceptable *precision and recall.* As we start experimenting with relative cut-off thresholds, we begin to see better results. Using thresholds based on cumulative probability (CP) cut-offs gives a better weight to the phrases in a category's context. A 0.1 cumulative probability gives a better weight across all categories as shown in Table 8 and Table 9

Table 8. Precision at various thresholds

Precision						
Category	Absolute	0.08CP	0.1CP	0.13CP	0.18CP	0.25CP
Grain	65.52%	75.38%	80.33%	72.46%	62.35%	64.63%
Crude	100.00%	100.00 %	98.53%	100.00%	100.00%	100.00%
Money-Fx	70.59%	77.59%	86.79%	69.84%	63.01%	67.14%
Ship	100.00%	73.47%	77.08%	82.98%	84.78%	86.36%
Metal	20.83%	54.05%	75.86%	78.72%	67.35%	81.08%
Micro	80.92%	84.14%	89.87%	86.21%	80.00%	84.34%
Macro	71.39%	76.10%	83.72%	80.80%	75.50%	79.84%

Table 9. Recall at various thresholds

Recall						
Category	Absolute	0.08CP	0.1CP	0.13CP	0.18CP	0.25CP
Grain	65.52%	84.48%	84.48%	86.21%	91.38%	91.38%
Crude	65.49%	72.12%	88.94%	79.65%	69.47%	80.53%
Money-Fx	76.60%	95.74%	97.87%	93.62%	97.87%	100.00%
Ship	3.45%	62.07%	63.79%	67.24%	67.24%	65.52%
Metal	8.33%	33.33%	36.67%	61.67%	55.00%	50.00%
Micro	51.00%	69.71%	79.06%	77.95%	73.05%	77.95%
Macro	43.88%	69.55%	74.35%	77.68%	76.19%	77.49%

We notice that as we start using cumulative probabilities as thresholds, we begin to see an improvement in *precision and recall.* The effectiveness of the classifier improves with relative thresholds. As the cumulative probability increases, the number of words characterizing a category increases too. We begin to see a bigger measure of typicality in each document. The adverse effect of this is the reduction in semantic distances between the documents. Thus we see in later experiments a larger number of documents being classified under "**No Cat**", as the hit rates are equal in more than one category.

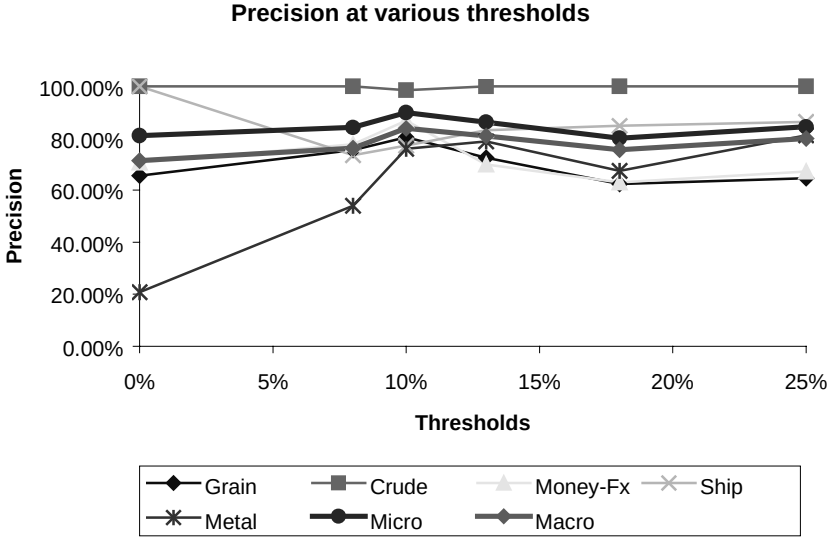

Fig. 2. Precision values at various thresholds

From Figure 2 and Figure 3 we also notice the values of *precision and recall* changing with the experiments. *Recall* seems to improve with increase in cumulative probability cut-off, whereas *precision* seems to work the other way. We notice also that an optimum value of *precision and recall* results at a cumulative probability of 0.1.

We notice the measurements are lower for the categories METAL and SHIP. This could be because of the smaller training set. The bigger the training set, the better the phrasal descriptors that can be extracted.

Another interesting inference drawn is that this classifier works well in the context of the Reuters-21578 corpus. As the phrases are drawn from this corpus, the jargons, acronyms and phrases local to Reuters are picked up as the descriptors. But a more generic classifier can be designed if the local expressions – jargons, acronyms, phrases, short-forms – are identified from this corpus and added to the stop list.

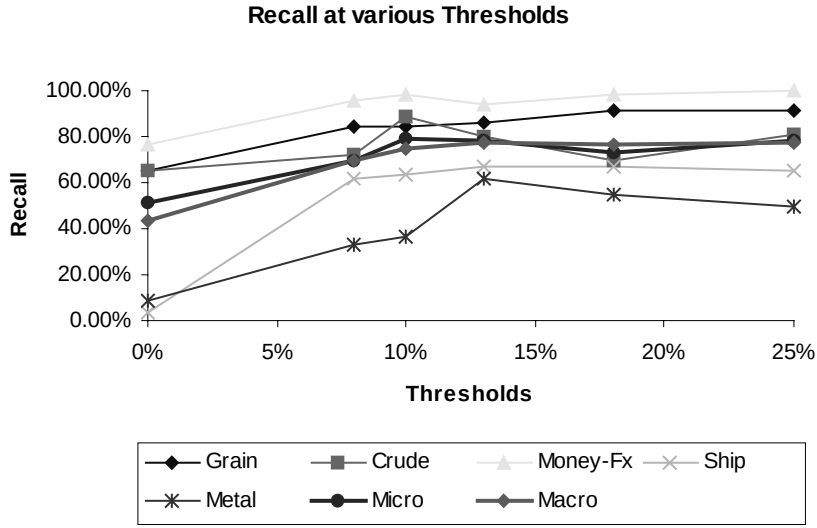

Fig. 3. Recall values at various thresholds

Suppose

T = Total number of test documents

N= Number of documents that cannot be categorized

Then effective decision ED = (T-N)/T

%ED = (T – N)/T * 100

This new measurement of effective decisions, as seen in Figure 4, seems to give a better perspective of the classifier's effectiveness. Again, we notice that all the graphs are flat after a cumulative probability threshold of 0.1.

Experiment with *Words*

To compare the phrase-based approach to a word-based approach, we conducted experiments using a similar classifier, with words constituting the category profile, instead of phrases.

Our initial attempts were not too successful, as certain words in most of the categories made up more than three percent of the total profile. For example, the word "vs" in the CRUDE category made up more than 5% of the total words that described the category. We had to eliminate this word. On the other hand, "vs" along with its adjoining word formed a meaningful phrase in the experiment with phrases. So we had to come up with more stop words to stem out these spurious words that did not contribute towards characterizing the category.

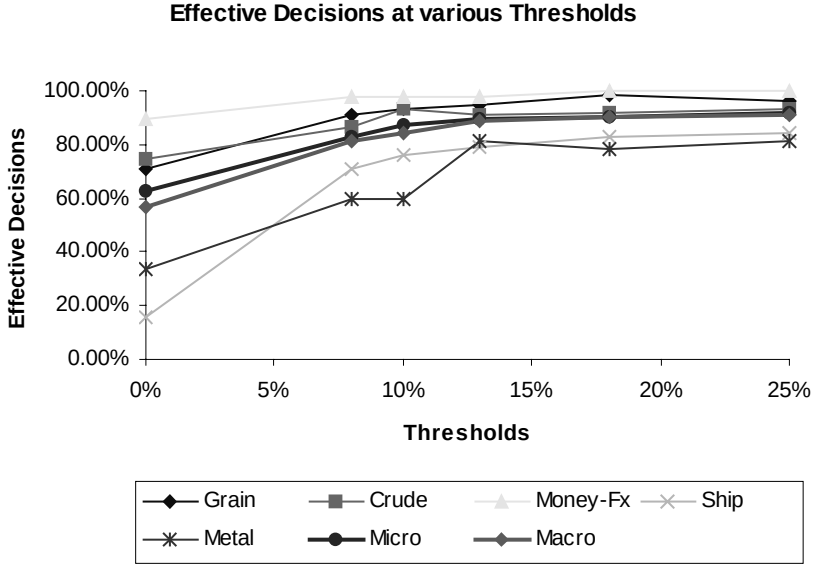

Fig. 4. Effective decisions at various thresholds

We conducted six experiments using the same thresholds as the ones we had used for the phrases.

For the first run, we used an absolute number of words across all the categories – 15 words. The results of the runs are shown in Figure 5.

The next five runs were based on relative thresholds of 0.08, 0.1, 0.13, 0.18 and 0.25 cumulative probabilities as the cut-off. The results of the runs from relative thresholds are shown in Figure 6 and Figure 7.

Though the results from the two runs – phrases-based and words-based – are not really comparable, as the stop words for the words-based experiment had to be increased to stem out high frequency low descriptive (of the category) words, it definitely can be studied closely to make some valid deductions. The results from the words-based experiment definitely seem to have better precision and recall with absolute thresholds. But with relative thresholds we see a better precision with phrases at lower thresholds and then the value holds steadily. Recall seems to peak with phrases with a relative threshold between 0.1 and 0.13CP, unlike with words where the recall steadily climbs up at higher thresholds.

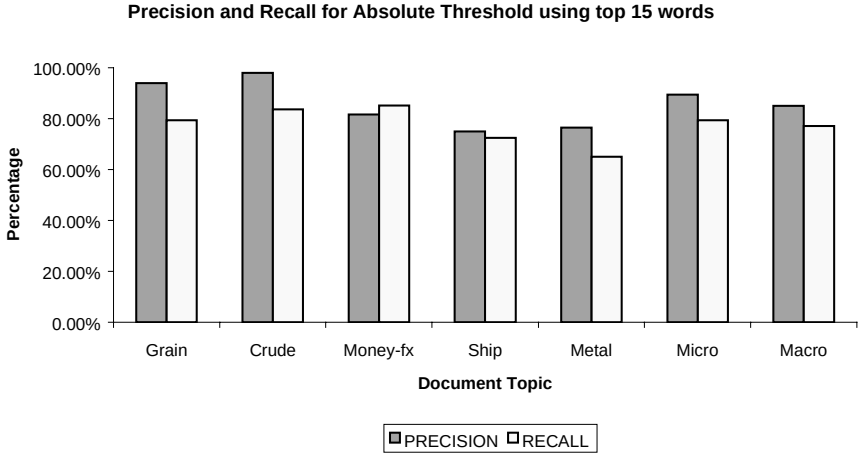

Fig. 5. Results of the words-based experiment using absolute threshold

5. Conclusions

Words and phrases are the salient features involved in classifying magazine articles. The number of different features in a corpus of magazine articles is enormous. Because the text classes regard the main topics of the texts, it is important to identify content terms that relate to the main topics and to discard terms that do not bear upon content or treat only marginal topics in training and test corpuses. The results of the empirical work conducted clearly emphasize the importance of phrases in classifying texts. Though our experiments with words yielded good results, we found instances and situations where the phrase-based approach produced more effectiveness. This could be due to the fact that when a word along with its adjoining word – a phrase – is considered towards building a category profile, it could be a good discriminator. This tight packaging of two words could bring in some semantic value. The packing of two words also filters out words occurring frequently in isolation that do not bear much weight towards characterization of the category.

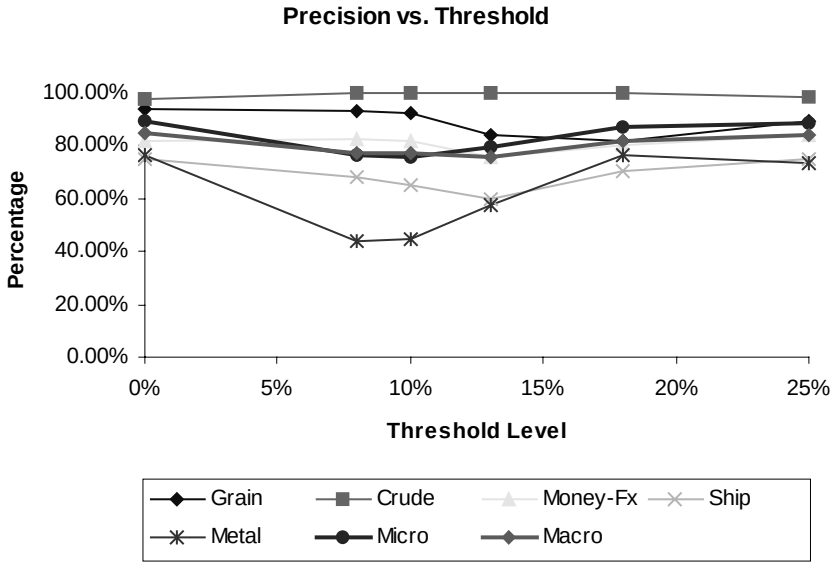

Fig. 6. Precision for the words-based experiment

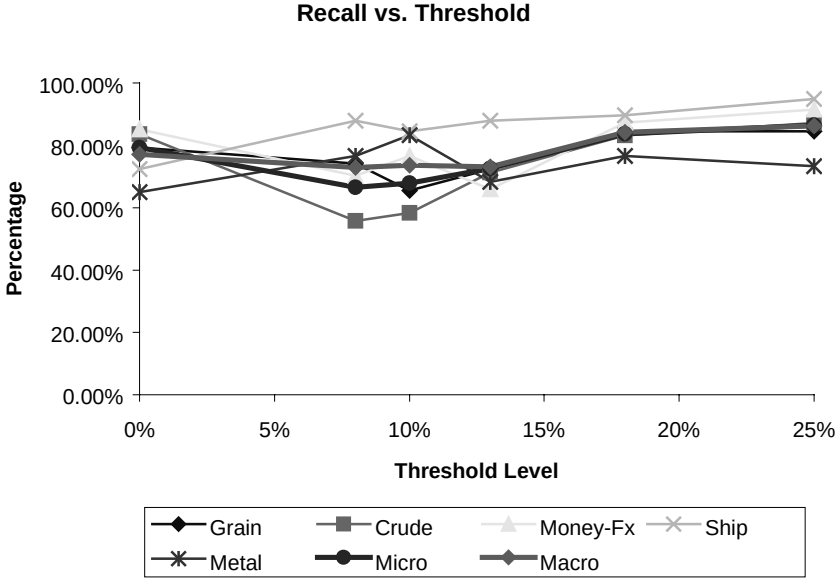

Fig. 7. Recall for the words-based experiment

Seemingly the graphs contradict this inference, as the macro recall and macro precision for the words are, at the least, as good as the ones for phrases if not better. The macro averages for the phrases have been weighed down by the categories SHIP and METAL. These categories' poor phrasal descriptors, resulting from small training sets, have lowered the measurements. Our conclusions are better substantiated by the graphs of GRAIN, CRUDE and MONEY-FX.

Our results seem more promising than the results of Lewis's[8] and Dasigi's[2] text categorization works based on the usage of phrases, for more than one reason. Both their works dealt with quite a big number of topics/categories employing techniques like *clustering*, *neural networks* and *Vector Space Modeling*, as opposed to our five categories of smaller test set and a straightforward method of category profile match-score. Another important distinction was that Dasigi's category profiles were entirely made up of single words. Their use of phrases was in the phrase-document matrix to which LSA [4] was applied. Our use of phrases is simply in the category profiles and therefore constitutes a somewhat different kind of use for phrases. We also found that our definition of a phrase gained better relevance in the context of Reuters-21578 corpus because the frequent use of phrases like "billion dlrs", "mln stg", "mln dlrs", "mln tonnes", "dlrs per", "mln acres", "mln vs", "cts vs", "shrs cts", "pct pay", etc, in the corpus helped us build a good set of category profiles for phrases rather than for words, with a better chance of profile matches. A complete list of phrases and words generated per topic as a part of this experiment is documented in the thesis report, which is at the Southern Polytechnic State University library.

The empirical research reveals a gray area of distinct classification – the ambiguity of a story falling into any one category because of high typicality values in more than one category. This always happens in the real world. Text classification by humans will always be subjective to a certain degree. We can try to refine the extraction or the classification process to reduce the ambiguity. The results of this experiment are definitely encouraging and pave the way for future research in the area of phrases based approach for text classification.

Limitations of the Research

As the experiments were conducted using just five categories and a small test set, the results – precision and recall – might appear more significant than they really are. To get more substantial results we need to expand our text corpus to comprise a higher – say at least 15 – number of categories with a much bigger set of test documents per category. The results of the empirical work can always be extrapolated to wider contexts.

Where Do We Go from Here?

Our goal in this research was to determine how meaningful it is to use phrases as discriminators, in categorizing documents. Although our experiment indicated

effectiveness for phrases, we still need to expand the scope of this research to establish conclusively that phrases are indeed effective.

Future researchers do not need to restrict themselves to just one data corpus. With more data sources, researchers can come up with a better word or phrase profiling, as the source will not be parochial. We realized in our analysis of phrases that the Reuters-21578 corpus does contain a lot of words, jargons, phrases, acronyms and a lot of other language inflections that are very familiar to their domain.

Our empirical experiment classified a big number of articles under “No Category”, as there was no clear-cut score for the article to fall under any one category. Future research work could focus on refining the algorithm to reduce this ambiguity of “No Category”. Another experiment that could be conducted is to follow a two-pass approach in classification, where the first pass involves classification by words and the second pass could further refine the experiment using phrases or synonyms.

References

1. Cañas, A.J., F. R. Safayeni, D. W. Conrath, *A Conceptual Model and Experiments on How People Classify and Retrieve Documents*. Department of Management Sciences, University of Waterloo, Ontario, Canada, 1985.
2. Dasigi. V, Mann C. Reinhold, Protopopescu A. Vladimir, “Information fusion for text classification – an experimental comparison”, in *The Journal of The Pattern Recognition Society*, 34(Sept 2001) 2413-2425.
3. Dasigi, V. and N. Verma: Automatic Generation of Category Profiles and their Evaluation through Text Classification, Proc.2nd International Conference on Intelligent Technologies, November, 2001, pp. 421-427.
4. Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer and Richard Harshman, "Indexing by latent semantic analysis", in *Journal of the American Society for Information Science*, 41(6), 391-407, 1990.
5. Sebestiani, Fabrizio. Attardi, Guiseppe , "Theseus: Categorization by context”, Giuseppe Attardi Dipartimento di Informatica Universit di Pisa, Italy...(1999).
6. Fuhr, Norbert, Stephen Hartman, Gerhard Lustig, Michael Schwanter, Konstadinos Tzeres and Gerhard Knorz, "Air/X-- a rule based multistage indexing system for large subject fields, In *RIAO 91 Conference Proceedings: Intelligent Text and Image Handling*, 606-623, 1991.
7. Lewis, David D., “Representation and Learning in Information Retrieval” Ph.D. thesis, Department of Computer Science; University of Massachusetts; Amherst, MA, 1992.
8. Lewis, David D., “An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task”, *Fifteenth Annual International Association for Computing Machinery SIGIR Conference on Research and Development in Information Retrieval*, Copenhagen, 1992, 37-50.
9. Ittner, D.D., Lewis, D.D., Ahn, D., “Text categorization of low quality images”. In *Proceedings of SDAIR-95, 4th Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, US, 1995, 301–315.
10. Moens, M.-F. and Dumortier, J., *Automatic Categorization of Magazine Articles*, Katholieke Universiteit Leuven, BelgiumInterdisciplinary Centre for Law & IT (ICRI).
11. Riloff, E., W. Lehnert, "Information Extraction as a Basis for High-Precision Text Classification," *ACM Transactions on Information Systems*, 12 (3), 1994, 296--333.

12. Rosch, E., "Principles of Categorization," in *Cognition and Categorization*, E. Rosch, B. B. Lloyd (Eds.), (Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1978), 27--48.
13. Smith, E.E., "Categorization," in *An invitation to Cognitive Science, Vol. 3, Thinking*, D. N. Osherson, E. E. Smith (Eds), The MIT Press, 1990, 33--53.
14. Yang, Y., An Evaluation of Statistical Approaches to Text Categorization, Technical Report CMU-CS-97-127, Computer Science Department, Carnegie Mellon University, 1999.