

Evaluating Text Categorization in the Presence of OCR Errors

Kazem Taghva, Tom Nartker, Julie Borsack, Steve Lumos,
Allen Condit and Ron Young

Information Science Research Institute
University of Nevada, Las Vegas
Las Vegas, NV 89154-4021

ABSTRACT

In this paper we describe experiments that investigate the effects of OCR errors on text categorization. In particular, we show that in our environment, OCR errors have no effect on categorization when we use a classifier based on the naive Bayes model. We also observe that dimensionality reduction techniques eliminate a large number of OCR errors and improve categorization results.

Keywords: OCR, errors, text categorization, naive Bayes model, categorization, classification

1. INTRODUCTION

One of the main objectives of information management is to develop tools to aid finding relevant documents to information requests. This objective becomes more prominent as the amount of information increases drastically. A natural and intuitive tool in information management is text categorization. The science of text categorization allows us to divide our entire document collection into a pre-defined set of thematic topics. Traditionally, the assignment of documents to these pre-defined categories were carried out by human experts. Recent advancements in technology, machine learning, and information retrieval is playing a very important role in automating the categorization task.

Currently, there are two distinct approaches to text categorization. The first approach is rule-based which is heavily dependent on machine learning algorithms and expert systems techniques. The second approach is a statistically based approach resembling the techniques of traditional information retrieval systems. In both these approaches, the systems are trained using a pre-assigned set of documents. Both approaches analyze these documents to identify *features* or *rules* to be used for categorizing future documents. An example of rule-based systems is CONSTRUE that uses machine learning algorithms to build a set of conditional clauses for each category.¹ Examples of statistically based systems are BOW² and ATTICS³ which are based on Bayes formulas. Since both text categorization approaches rely on the text of the documents for the purpose of training and category assignment of new documents, an interesting problem may be to investigate the effects OCR errors may have on categorization.

In this paper, we report on whether OCR errors have any effect on training categories or on the categorization of documents in a Bayesian text categorization system. The paper is organized as follows: Section 2 provides the theoretical background for the Bayesian system we use, section 3 gives a general description about our experiment, section 4 gives the results, and section 5 provides our conclusion and future work.

2. PROBABILISTIC CLASSIFIERS

Most of the statistically based text categorization techniques are based on the probabilistic approach introduced by Maron.^{4,5,6} In our experiments, we use the text categorizer BOW² which is based on the multinomial naive Bayes model.⁷ Following McCallum and Nigam,⁷ assume we have a vocabulary $V = (w_1, w_2, \dots, w_{|V|})$ for our collection, then a document d_i can be represented by a vector:

$$d_i = (N_{i1}, N_{i2}, \dots, N_{i|V|}) \quad (1)$$

where N_{ij} is the number of occurrences of the word w_j in the document d_i . We also assume we have a set of $C = \{c_1, c_2, \dots, c_{|C|}\}$ classes that we want to assign to our document collection. One basic assumption is that each document falls into exactly one category (i.e. exhaustive and incompatible).

In this framework, we are interested in finding $P(c_j|d_i)$, or the conditional probability that a document belongs to category c_j . Using Bayes' theorem, we can calculate this probability by:

$$P(c_j|d_i) = P(c_j) * \frac{P(d_i|c_j)}{P(d_i)} \quad (2)$$

In other words, Bayes' theorem provides a method to compute $P(c_j|d_i)$ by estimating the conditional probability of seeing particular documents of class c_j and the unconditional probability of seeing a document of each class. If we make the *word independence* assumption which states that the probability of each word occurring in a document is independent of the occurrences of other words in the document, then this probability can be estimated by:

$$P(d_i|c_j) = P(|d_i|)|d_i|! \prod_{t=1}^{|V|} \frac{P(w_t|c_j)^{N_{it}}}{N_{it}!} \quad (3)$$

In this formula, the $P(w_t|c_j)$ is computed using word frequencies in training documents. Assuming a training set of documents $D = \{d_1, d_2, \dots, d_{|D|}\}$ and the fact that we have an exhaustive and incompatible set of classes, then

$$P(W_t|c_j) = \frac{1 + \sum_{i=1}^{|D|} N_{it}P(c_j|d_i)}{|V| + \sum_{s=1}^{|V|} \sum_{i=1}^{|D|} N_{is}P(c_j|d_i)} \quad (4)$$

In general, various dimensionality reduction techniques are used to decrease the size of the vocabulary, increase the efficiency of the computation, and avoid “over-fitting.” These dimensionality reduction can be done locally or globally. In the local dimensionality reduction, a set of words is selected specific to each category. In this case, each category c_j has a different dimension.⁸ In the global dimensionality reduction, a set of words are chosen to represent all categories. One simple technique in decreasing the size of the vocabulary is by removing all words which occur in n or fewer document with n ranging between 1 and 10. Another variation of this approach is to remove all words with frequency less than or equal to n . Both these approaches have been used extensively with good results.^{9,10} More sophisticated approaches are based on the information content of words using entropy.^{11,12} In our experiments, we employ three of these techniques. These are described in Section 3.

3. EXPERIMENTAL ENVIRONMENT

Experimentation of automated text categorization systems has been ongoing since the mid seventies to improve and compare various algorithms. Our experiments though, are driven by the prevailing requirements of the Nuclear Regulatory Commission's (NRC) Licensing Support Network (LSN). Regulatory guidelines mandate that certain relevant federal records must be included to support the licensing process. The regulatory guidelines that define the inclusionary documents are topical in nature and fit nicely into the framework of text classification.

One conspicuous difference between the NRC's data and other experimental test collections is that the text of the NRC documents is optically recognized. We've shown in several IR experiments that in general, OCR text has no effect on retrievability.^{13,14} Does this hold true for text categorization as well? This was an important question that the NRC needed answered.

The NRC Collection is not static. As new information pertaining to the Repository at Yucca Mountain becomes available, these documents will need to be categorized appropriately. Currently, the collection contains in excess of 988,000 optically recognized documents. Over the next several years, this number will increase ten fold.

The NRC collection consists of page images as well as OCR generated text. The documents range from one page “Deliverable Acceptance” forms to government reports that are thousands of pages in length. Statistics for the collection used in our experimentation appear in Table 1.

Collection Statistics	
Document count	400
Number of pages	42,499
Average document length (pages)	106.25
Median document length (pages)	31
Approximate OCR error rate (words)	14%

Table 1. Experimental collection statistics

Class	Description
2.2	The Natural Systems of the Geologic Setting: Climatological and Meteorological Systems
3.3	Geologic Repository Operations Area (GROA) - Underground Facilities
5.1	Overall System Performance Assessment: Basic Approach
12.2	Information for Preparation of a Geologic Repository Environmental Impact Statement: Socioeconomic

Table 2. Sample topical guidelines

*Regulatory Guide 3.69*¹⁵ provides a list of topics to aid LSN participants decide whether documentary materials should be submitted to the LSN. The topics were written purposely at a very high level so that participants would err on the side of inclusion rather than exclusion. For our categorization study, we employed 52 classes of the top-most levels of the hierarchical Regulatory Guide and two exclusionary categories. In all, there are 72 top level and subordinate categories. Sample classes appear in Table 2.

The classes are not disjoint. Documents can be classified into more than one class or, as noted above, into one of two exclusionary classes. Unfortunately, there were very few documents categorized for us when we began our experiment. Since, we have trained several Geology students in the classification process. In this experiment there are a total of 400 documents categorized. The number of documents assigned to each class is listed in Table 3.

Class	Count	Class	Count	Class	Count
00.1	30	03.0	2	07.3	2
00.2	2	03.1	13	07.4	2
01.0	26	03.2	3	08.0	11
01.1	8	03.3	24	08.1	5
01.2	8	03.4	4	08.2	3
01.3	9	03.5	5	08.4	3
01.4	4	03.6	2	09.0	7
01.5	3	04.1	9	10.0	3
01.6	24	04.2	6	10.1	5
01.8	5	04.4	11	10.2	2
02.0	3	05.0	2	10.3	2
02.1	7	05.1	17	10.4	2
02.2	12	05.2	20	10.5	1
02.3	6	05.3	4	10.6	2
02.4	13	05.4	2	12.0	47
02.5	7	06.0	1	12.1	29
02.6	5	07.1	5	12.2	10
02.7	5	07.2	4	12.3	14

Table 3. Number of documents assigned to each class

Class	Default	Document Count	Occurrence Count	Information Gain
00.1	30.00	86.67	86.67	83.33
02.4	69.23	69.23	69.23	69.23
03.3	62.50	79.17	79.17	79.17
05.1	70.59	70.59	70.59	70.59
12.2	70.00	70.00	70.00	70.00
05.2	35.00	40.00	40.00	40.00

Table 4. Accuracy rate for each dimensionality reduction

Selection Method	# of Features	% of Misspellings
Default	102,060	73%
Document Count	22,286	23%
Occurrence Count	18,209	27%
Information Gain	10,000	21%

Table 5. Percentage of misspellings for each dimensionality reduction

Our experiments were run using BOW, the multinomial classification system, described in Section 2. Because of the small number of classifications available, we use the “leave one out” training method. We apply the four available dimensionality reduction techniques:

Default: No pruning of the vocabulary.

Document Count: Remove words that occur in N or fewer documents, $N = 3$.

Occurrence Count: Remove words that occur less than N times $N = 10$.

Information Gain: Remove all but the top N words by selecting words with the highest information gain $N = 10,000$.

An analysis of our results follows in Section 4.

4. EXPERIMENTS AND RESULTS

As we have stated, the purpose of our experimentation is to determine the effect of OCR errors on text categorization. The *accuracy rate* of a class is the ratio of the *number of correct decisions* over the total number of documents in the class. After running the experiments on the 54 classes with the reduction techniques listed in Section 3, we looked at six of the categories more closely to examine the effects of misrecognized words on document classification. We selected classes that had a fair number of training documents; five had good accuracy rates and one had done poorly. Table 4 shows the results for the **Default** run, which applies the complete index for classification, and each of the dimensionality reduction runs.

Note that each of the dimensionality reduction runs for each of the classes we reviewed improved or remained the same with respect to the **Default** run. Forty-six of the fifty-four classes showed improvement over the default. Knowing that most OCR errors are rarely duplicated (except under some unusual conditions),¹⁶ it is reasonable to assume that by pruning the vocabulary, most, if not all OCR errors would be eliminated from the training sets. The list of terms for each run was spell-checked using a domain specific dictionary. The percentage of misspellings appears in Table 5.

Of course, it has been shown in other experiments⁷ that dimensionality reduction, without the complication of OCR errors, tends to improve classification. So eliminating extraneous terms, misrecognized or just not useful, aids classification. But we believe that in particular, dimensionality reduction is an essential step when OCR text makes up the collection.

The general outline for the Retrieval Strategy Report is described below:

Cover Pages

0

These pages contain the complete title, document identifier, WBS number, SCPB identification number, QA designation, and the signatures and dates for the individuals who prepare, approve and concur with the approval of the document.

Executive Summary

This section provides a top level description of the study objective, findings, and recommendations.

Table of Contents

Figure 1. Poorly recognized OCR text

Selection Method	% of Misspellings
Default	39%
Document Count	16%
Occurrence Count	13%
Information Gain	7%

Table 6. Percentage of erroneous terms for document MOL.19971009.0431

For these classes, we also examined each of the documents that did not get categorized correctly. There was only one document (MOL.19971009.0431) with OCR text that we would consider to be poor. A paragraph of this document appears in Figure 1. For the terms that were selected for this document in the **Default** run, 39% were flagged as misspellings. The percentage of erroneous terms is greatly reduced for each of the dimensionality reduction runs. Table 6 shows the drop in percentages for this document. This table shows that by applying dimensionality reduction using BOW's global methods, OCR errors are inherently removed.

We also manually corrected all the errors in document MOL.19971009.0431. In all, 598 corrections were made. We reran our test and even with every OCR error corrected, the document was still not properly classified.

Note that not *all* these "misspellings" are OCR errors. Several may be proper nouns, acronyms, or misspellings that occurred in the original document. But with respect to each other, we believe these percentages are meaningful. Of course, the **Default** run has a much higher percentage of misspellings — nearly three fourths. But also, compare the lower percentage of misspellings in **Document Count** to **Occurrence Count**. This could be due to the fact that document quality can result in repetitive OCR errors in a single document, i.e., the same character in the same word is consistently misrecognized.

After examining the terms in the reduced feature sets, we noted that there were still OCR errors. Without corresponding corrected text, we could not definitively state that using OCR text makes no difference in text categorization. We corrected the OCR errors identified in the **Information Gain** run and reran the categorization for this purpose.

Of the 2152 misspellings in the **Information Gain** run, we made 1668 corrections. Nearly 80% of the terms flagged as "misspellings" were erroneous. There were 69,162 corrections made in the collection. The new **Information Gain** feature set differed from the original by 1158 words. Of these, 248 were still flagged as misspellings. The results of the original run and the corrected run appear in Table 7.

There was virtually no change in the accuracy of the classes. In fact, class 03.3 dropped slightly after the corrections were made. Our experiments show that OCR errors have little effect on text categorization once some form of dimensionality reduction has been applied.

Class	Info Gain Original Accuracy Rate	Info Gain Corrected Accuracy Rate
00.1	83.33	83.33
02.4	69.23	69.23
03.3	79.17	75.00
05.1	70.59	70.59
12.2	70.00	70.00
05.2	40.00	40.00

Table 7. Original and corrected Information Gain runs

We would like to note however that our collection size is small, only 400 documents. We only analyzed six of the 54 classes in any detail. But based on this study and previous IR and OCR studies, we believe that our results will carry forward to larger text categorization experiments.

5. CONCLUSION AND FUTURE WORK

There are two flavors of probabilistic models with the naive Bayes assumption in the literature. The one used in our experiment is called multinomial which takes word frequencies and document length into consideration to compute $P(c_j|d_i)$. This model typically works well with longer documents and large vocabulary sizes in the range of 10 to 20 thousand words. The second flavor which was originally proposed by Maron^{4,5} known as multi-variate only considers the presence or absence of the word in the document for computation of $P(c_j|d_i)$. Typically, the multi-variate model works best with shorter documents and if the size of vocabulary is limited to a few hundred words.⁷

Our experiment with multinomial naive Bayes classifiers implies that OCR errors have no effect on text categorization. Since the collection we used is a small collection of long documents, we believe the same experiment on a larger collection of varying document size may shed further light on this result. It will also be interesting to know if the same result holds on shorter documents using the multi-variate model. Another set of experiments could be conducted for rule based classifiers.

REFERENCES

1. P. J. Hayes, P. M. Andersen, I. B. Nirenburg, and L. M. Schmandt, "TCS: a shell for content-based text categorization," in *Proc. of CAIA-90, 6th IEEE Conf. on Artificial Intelligence Applications*, pp. 320–326, (Santa Barbara, CA), 1990.
2. A. McCallum, "Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering." <http://www.cs.cmu.edu/~mccallum/bow>, 1996.
3. D. D. Lewis, D. L. Stern, and A. Singhal, "ATTICS: a software platform for online text classification," in *Proc. of SIGIR-99, 22nd ACM Intl. Conf. on Research and Development in Information Retrieval*, pp. 267–268, (Berkeley, CA), 1999.
4. M. E. Maron and J. L. Kuhns, "On relevance, probabilistic indexing and information retrieval," *Journal of the ACM* **7**(3), pp. 216–244, 1960.
5. M. E. Maron, "Automatic indexing: An experimental inquiry," *Journal of the ACM* **8**, pp. 404–417, 1961.
6. D. D. Lewis, "Naive (bayes) at forty: The independence assumption in information retrieval," in *Proc. of ECML-98, 10th European Conf. on Machine Learning*, pp. 4–15, (Chemnitz, Germany), 1998.
7. A. McCallum and K. Nigam, "A comparison of event models for naive bayes text classification," in *AAAI-98 Workshop on Learning for Text Categorization*, 1998.
8. D. D. Lewis and M. Ringuette, "A comparison of two learning algorithms for text categorization," in *Proc. of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pp. 81–93, (Las Vegas, NV), 1994.
9. D. J. Ittner, D. D. Lewis, and D. D. Ahn, "Text categorization of low quality images," in *Proc. of SDAIR-95, 4th Annual Symposium on Document Analysis and Information Retrieval*, pp. 301–315, (Las Vegas, NV), 1995.
10. T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Proc. of ECML-98, 10th European Conf. on Machine Learning*, pp. 137–142, (Chemnitz, DE), 1998.

11. D. D. Lewis, "An evaluation of phrasal and clustered representations on a text categorization task," in *Proc. of SIGIR-92, 15th ACM Intl. Conf. on Research and Development in Information Retrieval*, pp. 37–50, (Kobenhavn, DK), 1992.
12. T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley, 1991.
13. K. Taghva, J. Borsack, and A. Condit, "Effects of OCR errors on ranking and feedback using the vector space model," *Inf. Proc. and Management* **32**(3), pp. 317–327, 1996.
14. K. Taghva, J. Borsack, and A. Condit, "Evaluation of model-based retrieval effectiveness with OCR text," *ACM Transactions on Information Systems* **14**, pp. 64–93, January 1996.
15. N. R. Commission, "Regulatory guide 3.69." <http://www.nrc.gov/NRC/RG/03/03-069.html>, 1996.
16. K. Taghva, J. Borsack, and A. Condit, "Results of applying probabilistic IR to OCR text," in *Proc. 17th Intl. ACM/SIGIR Conf. on Research and Development in Information Retrieval*, pp. 202–211, (Dublin, Ireland), July 1994.