Summarization as Feature Selection for Text Categorization

Aleksander Kołcz
Personalogy, Inc.
24 South Weber, Ste. 325
Colorado Springs, CO 80903
ark@personalogy.net

Vidya Prabakarmurthi
Department of Computer Science, University of Colorado at Colorado Springs
1420 Austin Bluffs Pkwy.
Colorado Springs, CO 80918
pravi@pcisys.net

Jugal Kalita
Department of Computer Science, University of Colorado at Colorado Springs
1420 Austin Bluffs Pkwy.
Colorado Springs, CO 80918
kalita@pikespeak.uccs.edu

ABSTRACT

We address the problem of evaluating the effectiveness of summarization techniques for the task of document categorization. It is argued that for a large class of automatic categorization algorithms, extraction-based document categorization can be viewed as a particular form of feature selection performed on the full text of the document and, in this context, its impact can be compared with state-of-the-art feature selection techniques especially devised to provide good categorization performance. Such a framework provides for a better assessment of the expected performance of a categorizer if the compression rate of the summarizer is known.

1. INTRODUCTION

There are several important criteria for judging the quality of a summarization technique, some of which are subjective by nature. From the perspective of a person it is important that a summary preserves the "gist" of a document. It is important that it can be used in place of the full document to judge the document's relevance to the current information need (e.g., while viewing the results of a Web search), or to otherwise categorize the document. Summaries, rather than full documents, are also used as inputs to machine learning systems. This is often due to the absence of full-text sources, but also because by using short summaries instead of complete documents the overall processing time can be substantially reduced.

One of the most important areas where summaries can be (and are) applied is categorization of text documents, where the goal of a system is to assign each document to one or more pre-defined categories. The exponential rate at which new documents become available creates significant challenges for systems that organize content for end users. Techniques that reduce the computational load involved are thus of increasing importance. For a large class of categorization algorithms, summarization (at least when it is based on extraction) can be viewed as a special kind of feature selection which dramatically shrinks the size of documents and, in turn, significantly reduces the number of features that need to be considered. Feature selection is not the primary focus of summarization algorithms, however, and it is important to understand how these algorithms compare in this respect with techniques specifically designed for that purpose.

Most studies focus on the subjective impact of summarization, where the quality of a summary and its utility as far as a particular task is concerned are judged by a group of human experts. In this work, we focus on the task of automatic document categorization in scenarios where a document's summary is functionally equivalent to reducing the number of features of the original. Generally, it is difficult to define an "ideal" summary for a document, although human experts are sometimes asked to do so. However, with a particular task in mind, it is possible to assess the impact of summarization on the performance of a system or a person, thus facilitating a comparison of competing techniques. We argue that, for the task of automatic text categorization, an ideal task-specific summary can be narrowly defined as the subset of most-informative features selected specifically with the categorization performance in mind. This results in an evaluation framework where the utility of any summarization technique can be assessed independently from others, and which makes it straightforward to specify the technique's *target* level of performance.

The paper is organized as follows. Section 2 reviews some of the prior work in this area. The feature-selection aspects of summarization are discussed in Section 3, whereas feature selection for text classifiers is addressed in Section 4. Section 5 reviews the traditional approaches to evaluating summarization systems and introduces the proposed framework. Experimental results are presented in Section 6 and the paper is concluded in Section 7.

2. RELATED WORK

Simple summarization-like techniques have been long applied to enrich the set of features used in text classification. For example, it has been common to give extra weight

CIKM'01, November 5-10, 2001, Atlanta, Georgia, USA.
Copyright 2001 ACM 1-58113-436-3/01/0011...$5.00.

to words appearing in the title of a story [19] or to treat the title words as separate features, even if the same words were present elsewhere in the text body [4]. It has been also appreciated that many documents contain useful formatting information, loosely defined as context, that can be utilized when selecting the salient words, phrases or sentences. For example, Web search engines weigh terms differently according to their HTML markup [2]. Summaries, rather than full documents, have been successfully applied to document clustering [6] and, recently, Ker and Chen [11] evaluated the performance of a categorization system using title-based summaries as document descriptors. In their experiments with a probabilistic TF-IDF based classifier, they showed that title-based document descriptors resulted in respectable levels of categorization performance.

3. SUMMARIZATION AS FEATURE SELECTION

Summarization techniques can be roughly divided into two groups: a) those baaed on abstraction of the original documents and b) those based on extraction from the original documents. The extraction-based approach imposes the constraint that a summary is formed by only using components of the original document (e.g., words, sentences or paragraphs [16]), while the abstraction-based approach leaves one relative freedom on how the summary is created. Although potentially more powerful, abstraction-based techniques have been far less popular than their exkaction-based counterparts, mainly because generation of the latter is more straight-forward (e.g., the leading paragraph of a document can be considered a simple extraction-based summary [5]). In this work, we focus on extraction-baaed methods exclusively.

Although summaries are specifically created to be read by people, they are often processed automatically as well. Depending on the type and method of processing it may or may not be important that a summary is syntactically valid and/or readable. Some of the most effective algorithms in text retrieval and classification are based on the "bag of words" representation where a document is treated as an unordered set of the terms. Here the positional information of a term is lost and its in-document frequency is often ignored as well (i.e., only the binary presence or absence of a term in a document is taken into account). It has been found that such a representation is sufficient for producing good results, e.g., when applied in conjunction with Naive Bayes [15], Support Vector Machines [4] or decision trees [1], particularly in text classification/categorization problems. Motivated by these findings, we consider only classifiers working with the binary bag of words representation.

An extraction-based summary consists of a sub-set of words from the original document and its bag of words representation can be created by selectively removing a number of features from the original term set. In text classification, such process is known as feature selection and is guided by the "usefulness" of individual features as far as the classification accuracy is concerned. In the context of text summarization, the feature selection aspect is only secondary. It might be even argued that in some cases a summary may contain the very same set of features as the original, for example, when it is created by removing only the redundant/repetitive words or phrases. Typically though, an extraction-baaed summary, whose length is only 10-15% of the original, is likely to lead to a significant feature reduction as well.

Many studies suggest that even simple summaries are quite effective in carrying over the relevant information about a document. From the document categorization perspective, their advantage over specialized feature selection methods (see Section 4) lies in their reliance on a single document only, the one that is being summarized, without having to compute the statistics for all documents sharing the same category label, or even for all documents in a collection. Moreover, various forms of summaries become ubiquitous on the Web and in certain cases their accessibility may grow faster than that of full documents.

4. FEATURE SELECTION FOR TEXT CLASSIFIERS

The problem of feature selection in text categorization has been researched in depth, and several very effective techniques have been identified [24]. Feature selection benefits a learner by eliminating non-informative or noisy features and by reducing the overall feature space to a manageable size. The latter factor is particularly important, since most learning algorithms suffer from the "curse of the dimensionality" and are unable to generalize well (at least without impractical amounts of training data) when the number of input parameters is too high. Numerous studies have shown that by reducing the feature space, the accuracy of a classification method can be increased and, even when only very few of the original features are kept, good accuracy can be maintained. Therefore, in principle, if each summary comprises some of the document's most informative features, the categorization accuracy obtained with summaries is likely to be high.

Most feature selection techniques use term frequency statistics within a category or across different categories and weigh terms according to their relevancy. If a separate classifier is built to distinguish each class from all others, one of the simplest approaches is to consider only the most frequent terms in that class. Perhaps surprisingly, such a basic technique is often quite effective [22], although it may identify insignificant terms (e.g., those equally common among different categories) while ignoring some of the more informative ones (e.g., terms that are relatively rare in one class but virtually absent from the others). Such inconsistencies can be overcome by accounting for the inter-class term statistics.

In particular, the Mutual Information (MI) and Chi-Squared (χ^2) criteria [24], which do consider term distribution among the classes, have been found to be particularly effective. Other techniques, such as term strength and odds ratio are largely comparable as well [18]. In our experiments, we used the Mutual Information (MI) criterion, which in a two-class setting can be defined as:

$$MI(t) = \sum_{t\in\{0,1\}} \sum_{c\in\{0,1\}} P(t,c) \log \frac{P(t,c)}{P(t)\,P(c)} \qquad (1)$$

where t denotes the term (i.e., feature), c is the class label and $P(t,c)$ is the joint probability of t and c cooccurring in a document collection; P (t) is the probability of a document (regardless of class) containing term t and $P(c)$ is the probability of a document belonging to class c. The probabilities in (1) can be usually estimated using simple maximum likelihood techniques (i.e., using empirical frequencies of terms and class labels in the training collection). Mutual Information assigns higher relevancy to features that are more common in one class than the other, which is quite intuitive. Since this feature selection technique can be considered state-of-the-art, we found it appropriate to use it as an "ideal" to which the effectiveness of a summarization technique as feature selector is compared. Note that there is certainly no single "best" feature selection method. However, the top performers such as MI or χ^2, tend to lead to largely equivalent results [24].

5. EVALUATION OF SUMMARIZATION SYSTEMS

5.1 Traditional Approaches

Evaluation of summarization systems [8] is inherently difficult due to multiple possible uses of a summary and the difficulty of defining an ideal one. Researchers have used both the *intrinsic* and extrinsic approaches, where an intrinsic evaluation typically selects a methodology of defining an ideal summary and then proceed to compare each summary with the corresponding ideal [5][3][9]. An extrinsic evaluation, on the other hand, bypasses the step of generating an ideal summary and instead captures the effectiveness of each summary at a certain task [8][20].

The recent large-scale TIPSTER[1] SUMMAC study [7] defined two such tasks: the ad-hoc task, in which summaries were used to judge the relevance of documents in the information retrieval context, and the categorization task, in which analysts were asked to categorize documents based on their summaries. The performance obtained with the original document collection (i.e., prior to summarization) is typically used as the baseline against which comparisons are made.

5.2 Feature Selection Approach

Traditional evaluation studies typically rely on human subjects, either for creating the ideal summaries, or for judging the usefulness of different summaries. We propose a hybrid approach specifically targeting evaluation of the performance of a summarization technique in automatic text categorization. In the process, we do define an ideal summary, but instead of measuring an explicit agreement of any given summary with the ideal, we compare the categorization performance obtained with the actual and ideal summaries. Arguably, the proposed evaluation methodology is quite narrow and ignores other important aspects of a summary. Therefore it should be applied to summarization techniques that have otherwise been judged "reasonable".

Let a summarization technique reduce the original document to N unique features. The task is to compare the categorization accuracy obtained with this feature set with the performance obtained using N "best" unique features extracted via a state-of-the-art feature selection technique. The N best unique features for a document, i.e., its "ideal summary", are obtained in the following manner:

- For all features available to the classifier (i.e., extracted from the training document collection) a relevance weight is assigned by the feature selection technique. This step is performed just once and its results are shared by all subsequent steps.
- For each document, its set of unique term features is identified and then ranked according to their relevance weights; the top N elements are retained to be used by a classifier.

Note that many categorization algorithms are sensitive to the number of features used, and sometimes using a richer feature set actually harms the performance. This poses certain difficulties in the standard setting, where the categorization performance using a summarized collection is compared with that using the original document collection. The current framework is immune to such dependencies since, regardless of the categorization technique, documents with comparable lengths are always used. Additionally, the proposed methodology makes it natural to pose (and answer) questions such as: *What* is *the* target level *of* performance for a summarization technique reducing a document size *to* 10% of *the* original? Also note that, once a particular state-of-the-art feature selector is chosen, the utility of any particular summarizer can be evaluated on its own, without the need for an extensive cross--method comparison.

6. EXPERIMENTAL SETUP

6.1 Document Corpus

We used the Reuters-21578 collection' as a testbed for our experiments. These documents represent short newswire stories which are perhaps not an ideal target for summarization. However, the Reuters dataset has been used extensively in text categorization studies and therefore has been adopted here. Although the complete dataset consists of 21,578 documents, we chose the popular *ModApte* split of the data [1], resulting in 9,603 training documents and 3,299 test documents. Additionally, we used only the 90 categories which have at least one training and one testing document and eliminated short documents for which summarization would not make sense (e.g., documents consisting of just the title). This resulted in 7,037 training documents and 2,734 test documents. Each article in the collection is formatted with SGML-like tags, thus providing for an easy identification of titles and paragraphs and facilitating the summarization process.

6.2 Categorization Method

From the large number of machine learning algorithms that are successful in text categorization we chose to use the Support Vector Machines (SVMs) [21][10] due to their high accuracy and an inherent ability to handle large feature spaces such as text. SVMs represent a relatively recent development in the area of statistical learning and belong to a class of algorithms that maximize the margin separating examples belonging to different classes in the high-dimensional

[1] http://www.tipster.org

[2] http://www.research.att.com/~lewis/reuters21578.html

input space. In particular, an SVM defines the classification boundary with only a subset of the original training set, known as the support vectors. In the linear case, the classification decision of an SVM takes the form of

$$SVM(x) = \text{sign}(w \; x - b)$$

where [·] denotes the dot-product, w and b are the weight vector and bias of a trained SVM, respectively, and x is the feature vector of an input document. If the training data are not linearly separable, an SVM transforms the original input space, via a nonlinear kernel transformation, into a higher dimensional one where the classification surface is more likely to have the form of a hyperplane. A detailed discussion of the SVMs and their associated training procedures can be found in [21].

The linear-kernel SVM was used in our experiments since it was previously demonstrated as very effective with the Reuters corpus [4][19][13]. We implemented the SVM classifier in the standard one-against-rest mode where, for a C-category problem, C separate classifiers were built, each distinguishing one category from all others.

6.3 Summarization Methods

To quantify the arguments advanced in this paper, we considered a number of simple extraction-based techniques — the details are given in list below. Similar heuristics-based techniques have been used for example in [16] [5] [17]. In all cases, a word occurring at least 3 times in the body of a document was considered a keyword, while a word occurring at least once in the title of a story was considered a *title word*. Common stopwords were removed from both the keyword and title-word sets.

- Title: the title of a story.
- FirstPara: the first paragraph of a story.
- ParaWithMostTitleWords: the paragraph which has highest title word count; if more than one exist, the first one counting from the top of the document — is chosen.
- ParaWithMostKeywords: the paragraph which has highest keyword count; if more than one exist, the first one counting from the top of the document — is chosen.
- FirstTwoPara: the first two paragraphs of a story.
- FirstLastPara: the first and the last paragraphs of a story.
- BestSentence: Summarizes by selecting (and maintaining their sequential order) those sentences in the story that contain at least 3 titlc words and at least 4 keywords.

The above methods are arguably simple but, for example, it was shown in [3] that lead-based summaries of news articles can be more informative than those resulting from more complex approaches. Also, headline-based article descriptors proved to be effective in determining users' interests [12].

Table 1: Average summary length and time (in seconds) to process the Reuters corpus for the summarization techniques considered. The time to perform MI-based feature selection is included as a reference.

summary	avg length	time
MI	*N/A*	320
Title	6	46
FirstPara	14	85
ParaWithMostTitleWords	15	180
ParaWithMostKeywords	16	935
FirstTwoPara	24	46
FirstLastPara	24	75
BestSentence	34	1246

Table 1 lists the average number of unique terms for each summarization technique, as well as the number of seconds used to compute the summaries for all documents in the dataset. By comparison, the MI-based feature extraction took 320 seconds, which is significantly longer than all methods with the exception of ParaWithMostKeywords and BestSentence. The measurements were performed using a system running Linux Red Hat v. 7.0, equipped with a 500 MHz Pentium-III microprocessor and 256 MBytes of RAM. The summarization and feature selection algorithms were coded in Perl v. 5.6.

6.4 Results

Only minimum preprocessing was applied to tokenize the documents/summaries prior to their use with the SVM categorizer, i.e., punctuation was removed, all characters were converted to lowercase and words were defined as sequences of consecutive characters separated by whitespace. For each of the summarization methods, two corresponding data sets were generated: one where, for each document, the MI feature selector picked the same number of unique terms as in the document's summary (see Section 5.2), and one where the average length of summaries was first computed and then applied uniformly by the feature selector. Categorization performance was measured by the micro-averaged breakeven point (BEP) [14] [23] between precision and recall. BEP, which is one of the most widely used categorization accuracy measures, is computed by merging the contingency tables of all one-against-rest classifiers and finding a decision threshold for which the numbers of positive and negative misclassification errors are equal, which translates to equal values of precision and recall. The BEP results obtained in our experiments are presented in Table 2. By comparison, the categorization performance using all features resulted in $BEP = 0.86$, which is consitent with the results published by others ([10][4][19]). It can be seen that, with the exception of ParaWithMostKeywords and FirstLastPara, the results due to summarization are comparable to those obtained with MI-based feature selection and, in some cases (most notably for the Title method), the features identified by the summarizer proved more effective than those selected by MI. It appears that, at least in this dataset, features located in the initial part of a document (including its headline) are most relevant for determining a document's category. This confirms the findings reported in other studies. Apart from the case of BestSentence, there was no difference between forcing MI to select exactly the same number

Table 2: Micro-averaged BEP results comparing the feature selection effectiveness of the summarization techniques (column SUM) against the baseline of using the Mutual Information feature selector (column MI); bold face points to cases where summarization was at least as good as standard feature selection. Column MI-AVG reports performance for a fixed-length MI feature selector (see text for details).

summary	SUM	MI	MI-AVG
Title	0.79	0.77	0.77
FirstPara	0.83	0.82	0.82
ParaWithMostTitleWords	0.82	0.82	0.82
ParaWithMostKeywords	0.78	0.82	0.82
FirstTwoPara	0.83	0.83	0.83
FirstLastPara	0.82	0.83	0.83
BestSentence	0.83	0.83	0.84

of features as in the summary and the case where MI selected the same number of features for all documents. In the case of BestSentence, the average summary length was quite high, and by applying it to all documents more highly informative features were selected for some documents, thus improving the overall performance.

7. CONCLUSIONS

We have proposed a framework for evaluating summarization methods in the context of their utility as feature selectors in automatic text categorization. Our approach is well suited for classifiers utilizing binary feature vectors, where a feature corresponds to the presence or absence of a word in a document. The advantage of the use of summarization in this context is that it is generic and can be performed much faster than many standard feature selection techniques, since it relies on inter-category term statistics.

The results obtained with the Reuters-21578 corpus demonstrate that (at least for this dataset) summarization can indeed be quite competitive with established feature selection techniques. The good performance can in part be attributed to the fact that news stories are written so as to capture users' attention with their headlines and initial text, which causes simple extraction-based summaries to contain highly relevant content.

8. ADDITIONAL AUTHORS

Additional authors: Joshua Alspector (Personalogy Inc., email: josh@personalogy.net).

9. REFERENCES

[1] C. Apté, F. Damerau, and S. M. Weiss. Automated learning of decision rules for text categorization. ACM Transactions on Information *Systems*, 12(3):233–251, 1994.

[2] R. K. Belew. Finding *out about:* A Cognitive *Perspective on Search* Engine *Technology* and *the WWW.* Cambridge University Press, 2000.

[3] R. Brandow, K. Mitze, and L. F. Rau. Automatic condensation of electronic publications by sentence selection. Information Processing and Management, 31(5):675–685, 1995.

[4] S. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorization. In Proceedings *of 7th* International Conference on Information and *Knowledge* Management, pages 229-237, 1998.

[5] H. P. Edmundson. New methods in automatic extracting. Technical report, Department of Computer Science, University of Maryland at College Park, 1969.

[6] V. Ganti, J. Gehrke, and R. Ramakrishnan. CACTUS - clustering categorical data using summaries. In *Knowledge Discovery* and Data Mining, pages 73-83, 1999.

[7] T. F. Hand and B. Sundheim. TIPSTER-SUMMAC summarization evaluation. In Proceedings *of the TIPSTER Text* Phase *III Workshop,* 1998.

[8] H. Jing, R. Barzilay, K. McKeown, and M. Elhadad. Summarization evaluation methods: Experiments and analysis. In *AAAI* Intelligent *Text* Summarization *Workshop,* pages 60–68, 1998.

[9] H. Jing and K. McKeown. The decomposition of human-written summary sentences. In Proceedings *of the 22nd* Annual International *ACM SIGIR* Conference on Research and Development in *Information Retrieval,* pages 60-68, 1999.

[10] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In Proceedings *of the Tenth* European Conference on *Machine* Learning *(ECML-98)*, pages 137-142, 1998.

[11] S. J. Ker and J. N. Chen. A text categorization based on a summarization technique. In *ACL '2000 Workshop on* Recent Advances in Natural Language *Processing and Information* Retrieval, 2000.

[12] A. Kolcz and J. Alspector. Asymmetric missing-data problems: overcoming the lack of negative data in preference ranking. *to* appear in Information *Retrieval,* 2001.

[13] J. T. Y. Kwok. Automated text categorization using support vector machine. In Proceedings *of the International* Conference on Neural Information Processing *(ICONIP),* pages 347-351, 1999.

[14] D. D. Lewis. Evaluating text categorization. In Proceedings *of Speech* and Natural Language *Workshop,* pages 312–318. Morgan Kaufmann, 1991.

[15] D. D. Lewis. Naive (Bayes) at forty: the independence assumption in information retrieval. In Proceedings *of the 10th* European Conference on Machine *Learning,* pages 4-15, 1998.

[16] II. P. Luhn. The automatic creation of literature abstracts. In *IRE* National Convention, pages 60-68, 1958.

[17] K. Mahesh. Hypertext summary extraction for fast document browsing. In *Working Notes of the AAAI* Spring Symposium on Natural Language Processing *for the World Wide Web,* pages 95-103, 1997.

[18] D. Mladenić and M. Grobelnik. Feature selection for classification based on text hierarchy. In Working *notes of* Learning from *Text* and *the Web:* Conference on Automatic Learning and *Discovery (CONALD-98)*, 1998.

[19] B. Raskutti, H. Ferrá, and A. Kowalczyk. Second order features for maximising text classification performance. In Proceedings of *the 12th* European Conference on Machine Learning, 2001.

[20] A. Tombros, M. Sanderson, and P. Gray. Advantages of query based summaries in information retrieval. In *Working Notes of the AAAI* Spring Symposium on Natural Language Processing *for the World Wide Web,* pages 44–52, 1998.

[21] V. N. Vapnik. Statistical Learning *Theory.* John Wiley, New York, 1998.

[22] S. M. Weiss, B. F. White, C. Apté, and F. Damerau. Lightweight document matching for help-desk applications. *IEEE Intelligent* Systems, 15(2), 2000.

[23] Y. Yang and X. Liu. A re-examination of text categorization methods. In Proceedings of *the* 22nd International *ACM SIGIR* Conference on Research and *Development in* Information Retrieval, pages 42-49, 1999.

[24] Y. Yang and J. P. Pedersen. A comparative study on feature selection in text categorization. In Proceedings *of the Fourteenth* International Conference on *Machine* Learning *(ICML'97)*, pages 412–420, 1997.