Summarization as Feature Selection for Document Categorization on Small Datasets

Emmanuel Anguiano-Hernández[1], Luis Villaseñor-Pineda[1], Manuel Montes-y-Gómez[1], and Paolo Rosso[2]

[1] Laboratory of Language Technologies, Department of Computational Sciences, National Institute of Astrophysics, Optics and Electronics (INAOE), Mexico
{eanguiano,villasen,mmontesg}@inaoep.mx
[2] Natural Language Engineering Lab, ELiRF, Department of Information Systems and Computation, Polytechnic University of Valencia (UPV), Spain
prosso@dsic.upv.es

Abstract. Most common feature selection techniques for document categorization are supervised and require lots of training data in order to accurately capture the descriptive and discriminative information from the defined categories. Considering that training sets are extremely small in many classification tasks, in this paper we explore the use of unsupervised extractive summarization as a feature selection technique for document categorization. Our experiments using training sets of different sizes indicate that text summarization is a competitive approach for feature selection, and show its appropriateness for situations having small training sets, where it could clearly outperform the traditional information gain technique.

Keywords: Text Categorization, Text Summarization, Feature Selection.

1 Introduction

Automatic document categorization is the task of assigning free text documents into a set of predefined categories or topics. Currently, most effective solutions for this task are based on the paradigm of supervised learning, where a classification model is learned from a given set of labeled examples called training set [7]. Within this paradigm, an important process is the identification of the set of features (words in the case of text categorization) more useful for the classification. This process, known as feature selection, tends to use statistical information from the training set in order to identify the features that better describe the objects of different categories and help discriminating among them. Due to the use of that statistical information, the larger the training set, the better the feature selection. Unfortunately, due to the high costs associated with data labeling, for many applications these datasets are very small. Because of this situation it is of great importance to search for alternative feature selection methods specially suited to deal with small training sets.

In order to tackle the above problem, in this paper we propose to apply unsupervised extractive summarization as a feature selection technique; in other words, we propose reducing the set of features by representing documents by means of a representative subset of their sentences. Our proposal is supported on two facts about

H. Loftsson, E. Rögnvaldsson, S. Helgadóttir (Eds.): IceTAL 2010, LNAI 6233, pp. 39–44, 2010.
© Springer-Verlag Berlin Heidelberg 2010

extractive summarization. First, it has demonstrated to capture the essence of texts by selecting their most important sections, and, consequently, a subset of words adequate for their description. Second, it is an inherently local process, where each document is reduced individually, bypassing the restrictions imposed by the size of the given training set.

Through experiments on a collection consisting of three training sets of different sizes we show that text summarization is a competitive approach for feature selection and, what is more relevant, that it is specially appropriate for situations having small training sets. Particularly, in this situations the proposed approach could significantly improved the results achieved by the information gain technique.

The rest of this document is organized as follows. Section 2 presents some related works concerning the use of text summarization in the task of document categorization. Section 3 describes the experimental platform; particularly it details the feature selection process and the used datasets. Then, Section 4 shows the results achieved by the proposed approach as well as some baseline results corresponding to the application of information gain as feature selection technique. Finally, Section 5 presents our conclusions and future work ideas.

2 Related Work

Some previous works have considered the application of text summarization in the task of document categorization. Even though these works have studied different aspects of this application, most of them have revealed, directly or indirectly, the potential of text summarization as a feature selection technique.

Some of these works have used text summarization (or its underlying ideas) to *improve the weighting of terms* and, thereby, the classification performance. For instance, Ker and Chen [2] weighted the terms by taking into account their frequency and position in the documents; whereas Ko et al. [3] considered a weighting scheme that rewards the terms from the phrases selected by a summarization method.

A more ambitious approach consists of applying text summarization with the aim of reducing the representation of documents and *enhancing the construction of the classification model*. Examples from this approach are the works by Mihalcea and Hassan [5] and Shen et al. [8]. The former work is of special relevance since it showed that significant improvements can be achieved by classifying extractive summaries rather than entire documents.

Finally, the work by Kolcz et al. [4] explicitly proposes the use of summarization *as a feature selection technique*. They applied different summarization methods –based on the selection of sentences with the most important concentration of keywords or title words– and compared the achieved results against those from a statistical feature selection technique, concluding that both approaches are comparable.

Different to these previous works, this paper aims to determine the usefulness of summarization as feature selection technique for the cases consisting of small training sets. Our assumption is that, because summarization is a local process, done document by document without considering information from the entire dataset, it may be particularly appropriate for these cases. Somehow, our intention is to extent the conclusions by Kolcz et al. by showing that, although summarization and statistical feature

selection techniques are comparable for most of the cases, the former is a better option for situations restricted by the non-availability of large training sets.

3 Experimental Platform

3.1 Feature Selection Process

Because of our interest to evaluate the effectiveness of text summarization as a feature selection technique, we compared its performance against the one of a traditional supervised (statistical) approach. Particularly, to summarize the documents we used the well-known $HITS_A$ *directed backward* graph-based sentence extraction algorithm [6]. The choice of this algorithm was motivated by its relevant results in text summarization as well as by its previous usage in the context of document categorization [5]. On the other hand, we considered the *information gain* (IG) measure as exemplar from supervised techniques [9].

In a few words, the feature selection was carried out as follows:

1. Summarize each document from the training set, by selecting the $k\%$ of their most relevant sentences, in line with the selected summarization method.
2. Define the features as the set of words extracted from the summaries, eliminating the stop words.

In contrast to this approach, the common (statistical) feature selection process defines the features as the set of words having positive information gain ($IG > 0$) within the entire dataset. That is, it selects the words whose presence or absence gives the larger information for category prediction.

3.2 Evaluation Datasets

For the experiments we used the *R8 collection* [1]. This collection is formed by the eight largest categories from the Reuters-21578 corpus, which documents belong to only one class. It contains 5189 training documents and 2075 test documents.

With the aim of demonstrating the appropriateness of the proposed approach for situations having small training sets, we constructed two smaller collections from the original R8 corpus: R8-41 and R8-10, consisting of 41 and 10 training documents per class respectively. These collections contain 328 and 80 training documents and the original 2075 test documents. Details can be found in Table 1.

Table 1. Documents distribution in different data sets

Class	R8 Training Set	R8-41 Training Set	R8-10 Training Set	Test Set
earn	2701	41	10	1040
acq	1515	41	10	661
trade	241	41	10	72
crude	231	41	10	112
money-fx	191	41	10	76
interest	171	41	10	73
ship	98	41	10	32
grain	41	41	10	9

4 Results

In the experiments we evaluated the effectiveness of feature selection by means of the classification performance. Our assumption is that, given a fixed test set and classification algorithm, the better the feature selection the higher the classification performance. In particular, in all experiments we used a Support Vector Machine as classification algorithm, term frequency as weighting scheme, and the classification accuracy and micro-averaged F_1 as evaluation measures.

Table 2 shows two baseline results. The first one corresponds to the usage of all words as features (i.e., without applying any feature selection method, except by the elimination of stop words); whereas, the second concerns the usage of only those words having positive information gain. From these results it is clear that the IG-based approach is pertinent for situations having enough training data, where it could improve the accuracy in 1.5%. However, it is also evident that it has severe limitations to deal with small training datasets. For instance, it only could define 20 relevant features for the R8-10 collection (which represented just 1% of the whole set of words), causing a decrement in the classification accuracy of around 50%.

Table 2. Baseline results: without feature selection and using the information gain criterion

	R8			R8-41			R8-10		
	Features	***Accuracy***	***F1***	***Features***	***Accuracy***	***F1***	***Features***	***Accuracy***	***F1***
All features	17,336	85.25	.842	5,404	78.75	.782	2,305	71.71	.702
IG > 0	1,691	86.51	.857	54	42.89	.539	20	35.57	.0402

Table 3 and table 4 show results from the proposed method for different summary sizes, ranging from 10% to 90% of the original sentences of the training documents. The achieved results are encouraging; they show that text summarization is a competitive approach for feature selection and, what is more relevant, that it is especially appropriate for situations having small training sets. In particular, for the reduced collections R8-41 and R8-10, very small summaries (from 10% to 50%) could outperform, with statistical significance, the results obtained by the application of the IG-based approach ($IG > 0$) as well as those obtained using all words as features. We evaluated statistical significance of the results using the z-test with a confidence of the 95%.

Table 3. Results accuracy of proposed method using summaries of different sizes

	R8			R8-41			R8-10		
Sum. Size	***Number features***	***Our method***	***Top IG***	***Number features***	***Our method***	***Top IG***	***Number features***	***Our method***	***Top IG***
10%	8,289	87.13	85.45	1,943	83.47	80.43	706	76.77	52.24
20%	9,701	88.53	85.54	2,445	82.27	78.02	902	70.17	56.87
30%	11,268	89.20	85.78	3,089	82.89	78.31	1,178	64.67	52.34
40%	12,486	87.90	85.78	3,569	83.52	78.60	1,392	75.23	54.07
50%	13,326	87.42	85.88	3,919	81.40	79.13	1,523	75.08	64.10
60%	14,560	86.89	85.64	4,348	79.66	78.89	1,722	69.40	67.52
70%	15,626	86.75	85.54	4,671	80.10	78.94	1,890	69.73	69.69
80%	16,339	86.70	85.69	5,004	80.43	78.31	2,082	71.23	69.83
90%	17,063	86.27	85.35	5,263	78.89	78.60	2,230	72.58	71.66

In order to have a deep understanding of the capacity of the proposed method, we compared its results against those from a classifier trained with the same number of features but corresponding to the top IG values (indicated in Table 4 as Top-IG). As can be noticed our method always obtain better results, indicating that the information gain cannot be properly evaluated from small training sets. Regarding this fact, it is interesting to notice that for the R8-10 collection, our method allowed a 7% of accuracy improvement (from 71.71 to 76.77) by means of a 70% feature reduction (from 2,305 to 706), whereas, for the same compression ratio, the features selected by their *IG* value caused a 28% drop in the accuracy (from 71.71 to 52.24).

Table 4. F1-measure of the proposed method using summaries of different sizes

	R8		**R8-41**		**R8-10**	
Sum. Size	***Our method***	***Top IG***	***Our method***	***Top IG***	***Our method***	***Top IG***
10%	.876	.846	.842	.817	.776	.572
20%	.886	.846	.834	.790	.709	.659
30%	.891	.848	.836	.789	.654	.618
40%	.877	.848	.842	.789	.766	.631
50%	.870	.848	.819	.791	.763	.700
60%	.864	.846	.798	.787	.683	.717
70%	.862	.845	.800	.786	.685	.716
80%	.861	.847	.803	.780	.693	.698
90%	.856	.843	.784	.781	.712	.703

5 Conclusions and Future Work

This paper studied the application of automatic text summarization as a feature selection technique in the task of document categorization. Experimental results in a collection having three training sets of different sizes indicated that summarization and information gain (a statistical feature selection approach) are comparable when there are enough training data (such as in the original R8 collection), whereas the former is a better option for situations restricted by the non-availability of large training sets (as in the cases of R8-41 and R8-10 collections). This behavior is because summarization is a local process, where each document is reduced individually without considering the rest of the documents; while statistical techniques such as IG require lots of training data in order to accurately capture the discriminative information from the defined categories. Due to this characteristic, as future work we plan to examine the pertinence of summarization-based feature selection into a semi-supervised text classification approach.

It is important to mention that the success of summarization depends on the nature of the documents. In this paper we evaluated the proposed method in a collection of news reports demonstrating its usefulness. As future work we plan to determine its appropriateness for other kinds of documents such as web pages and emails.

Acknowledgments. This work was done under partial support of CONACYT (Project grants 83459, 82050, 106013, and scholarship 271106), and the MICINN TIN2009-13391-C04-03 (Plan I+D+i) TEXT-ENTERPRISE 2.0 research project.

References

1. Cardoso-Cachopo, A.: Improving Methods for Single-Label Text Categorization. PhD Thesis. Technical University of Lisboa, Portugal (October 2007)
2. Ker, S.J., Chen, J.N.: A Text Categorization Based on a Summarization Technique. In: ACL 2000 Workshop on Recent Advances in Natural Language Processing, Honk Kong (2000)
3. Ko, Y., Park, J., Seo, J.: Improving Text Categorization Using the Importance of Sentences. Information Processing and Management (40) (2004)
4. Kolcz, A., Prabakarmurthi, V., Jugal, K.: Summarization as a Feature Selection for Text Categorization. In: Tenth International Conference on Information and Knowledge Management (CIKM 2001), Atlanta GA, USA (2001)
5. Mihalcea, R., Hassan, S.: Using the Essence of Texts to Improve Document Classification. In: RANLP 2005, Borovetz, Bulgaria (2005)
6. Mihalcea, R.: Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization. In: ACL 2004, Barcelona, Spain (2004)
7. Sebastiani, F.: Machine Learning in Automated Text Categorization. ACM Computing Survey 34(1), 1–47 (2002)
8. Shen, D., Yang, Q., Chen, Z.: Noise Reduction through Summarization for Web-Page Classification. Information Processing and Management (43) (2007)
9. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: Proceedings of ICML 1997, pp. 412–420 (1997)