IDIAP RESEARCH REPORT

Noisy Text Categorization

Alessandro Vinciarelli [a]

IDIAP–RR 03-61

November 2003

submitted for publication

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 − 27 − 721 77 11
fax +41 − 27 − 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[a] IDIAP - E-mail:vincia@idiap.ch

IDIAP Research Report 03-61

Noisy Text Categorization

Alessandro Vinciarelli

November 2003

submitted for publication

Abstract. This work presents a system for the categorization of noisy texts. By noisy it is meant any text obtained through an extraction process (affected by errors) from media different than digital texts. We show that, even with an average Word Error Rate of around 50%, the categorization performance loss with respect to the clean version of the same documents is negligible.

1 Introduction

Several media contain textual information that can be accessed only through an extraction process. Important examples are speech recordings and handwritten documents (that can be converted into text through a recognition process) or pictures and videos (where text can be detected and recognized). The common denominator of all above mentioned sources (the list is not exhaustive) is that the extraction process produces *noise*, i.e. word substitutions, deletions and insertions with respect to the actual *clean* text contained in the original source.

In the last decades, there have been many research efforts towards the modeling of textual information. Algorithms developed in domains like Information Retrieval allow the management and the effective use of huge document databases, but they have been applied essentially on clean texts. The possibility of extending such domains to noisy data would allow the management of databases of media different than digital text. While applications of Information Retrieval to noisy texts have been extensively studied (especially for what concerns speech recordings), no effort has been made, to our knowledge, to perform noisy text categorization. Text Categorization (TC) is the task of assigning a document one or more categories belonging to a predefined set $C = \{c_1, c_2, \ldots, c_{|C|}\}$, where $|C|$ is the cardinality of C [2]. The most effective approaches to the problem convert the documents into vectors and then apply machine learning techniques to perform the categorization. Among the best categorization algorithms, there are several models that need to be trained (e.g. Neural Networks and Support Vector Machines [2]). This makes the noise a problem because it can have rather negative effect on the training. Moreover, the collection of enough material to train good category models can be difficult for certain media.

A possible solution for both above problems is to train category models on clean texts (that are relatively easy to collect) and then apply them over noisy documents. The absence of noise in the training material improves the category models, but the mismatch due to noise between training and test conditions is likely to degrade their performance. This work shows that the performance loss when passing from clean to noisy versions of the same texts is negligible even when the average Word Error Rate (WER) is around 50%. In order to explain this result, we propose an information theoretic performance measure for the extraction process. The proposed measure accounts for the amount of information useful to the categorization preserved through the extraction process. This allows one to relate more clearly the text extraction and the text categorization performances.

The rest of this paper is organized as follows: Section 2 presents the TC algorithms we apply, Section 3 shows experiments and results and Section 4 draws some conclusions.

2 Text Categorization

One of the most effective TC approaches developed until now converts the documents into vectors and then categorizes them with Support Vector Machines [2]. The TC process used in this work is based on such approach and it is composed of several steps. First, the documents are preprocessed: all the characters supposed to be category independent (punctuation marks, digits, parentheses, etc.) are removed. At the end of this step, the documents are converted into streams of words and can be normalized. *Normalization* removes the variability that is not necessary in the TC task and it is composed of two steps: *stopping* and *stemming*. Stopping is the removal of all words that are assumed to be category neutral: articles, prepositions, words of common use (the so-called *stopwords*). Stemming is the replacement of inflected forms of a certain word (e.g. *connected*, *connection* and *connecting*) with their common stem (*connect*).

After preprocessing and normalization, the documents are available as streams of terms. This is not a suitable representation for the categorization and an *indexing* step is required to convert the documents into vectors where each component accounts for a term of the dictionary (the list of all terms appearing in the document database). The vector components are typically functions of term frequencies in the database. The most commonly applied function is the so called *tf.idf* [1]:

$$d_j = tf(j) \cdot \log\left(\frac{N}{N_j}\right) \tag{1}$$

where d_j is the component related to term j in document d, $tf(j)$ is the number of times term j appears in document d, N is the number of documents in the database and N_j is the number of documents containing term j.
Once the documents have been converted into vectors, it is possible to train SVM's performing the categorization. A different SVM is trained for each category $c \in C$. The SVM related to category c ranks the documents according to a score such that the higher the position in the ranking, the higher the confidence that the document belongs to c.

3 Experiments and Results

This section presents the experiments performed in this work. Categorization has been performed on both clean and noisy version of 100 documents and the performance difference is measured. The rest of this section is organized as follows: Section 3.1 describes the data, Section 3.2 shows the categorization results and Section 3.3 presents the text extraction performance measure we propose.

3.1 The Data

The experiments are based on the Reuters-21578 corpus, a well-known and widely used benchmark in TC research [2]. The corpus contains 12,902 documents split into training (9,603 documents) and test (3,299 documents) set (ModApté split). The corpus is composed of digital texts extracted from the Reuters Newswire Bulletin. The total number of represented categories is 90, but not all of them are well represented. The ten most represented categories account for more than 90% of the corpus, while many categories have less than 5 documents. For this reason, the experiments take often into account only the most frequent categories and this is the approach followed in this work.
We randomly selected five categories among the ten most frequent ones (*acquisitions*, *earnings*, *money-fx*, *grain* and *crude*). For each of them, a certain number of documents (30 for acquisitions and earnings and 20 for the others) were selected randomly from the Reuters test set. In order to obtain a noisy version, such documents have been manually written by a single person and then automatically transcribed with a handwriting recognition system [3]. This is not restrictive because the TC system does not take into account the original source of the noisy text. The noise measures do not depend on the media the text is extracted from and all the considerations that will be made in the following apply to noisy texts independently of the source they are extracted from.
The database of handwritten documents has been split into training (10 documents of the acquisitions category and 10 documents of the earnings category) and test set (20 documents of each category). Since the handwriting models are independent of the category, the fact that only two categories are represented in the training set is not a problem. The recognition system is based on HMM's and Statistical Language Models (trigrams) [3]. The recognition results over the test set are reported in Table 1 in terms of WER (percentage of misrecognized words) and Term Error Rate (the WER after the application of stopping and stemming).

3.2 Categorization Results

An SVM has been trained for each of the five categories using the training set of the Reuters database. The SVM's have a Gaussian kernel and the variance of the Gaussian has been set through crossvalidation [2]. Each SVM is a binary classifier trained to detect whether a document belongs or not to a certain category. Given a category c, $R(c)$ is the set of documents actually belonging to c and $R^*(c)$

Category	WER	TER
acquisitions	37.7%	26.4%
crude	52.5%	55.1%
earnings	76.4%	49.4%
grain	52.1%	44.6%
money-fx	42.7%	34.1%
average	52.3%	41.9%

Table 1: Handwriting recognition performance.

is the set of documents identified as belonging to c by the system. TC systems are evaluated by using Precision:

$$\pi(c) = \frac{|R^*(c) \cap R(c)|}{|R^*(c)|} \tag{2}$$

and Recall:

$$\rho(c) = \frac{|R^*(c) \cap R(c)|}{|R(c)|} \tag{3}$$

An overall measure of the performance can be obtained by averaging over all categories: $\pi = |C|^{-1}\sum_{c' \in C} \pi(c')$ and $\rho = |C|^{-1}\sum_{c' \in C} \rho(c')$. Precision can be interpreted as the probability that a document identified as belonging to c by the system actually belongs to c. Recall can be interpreted as the probability that a document actually belonging to c is identified as such by the system. The performance is typically measured by considering the π value at standard ρ levels (10%,20%,...,100%). This results in Precision vs Recall curves representative of the system performance.
Since the test set contains relatively few documents the results could be simply the effect of a random drawing. The ranking of N documents, where only L belong to a given category, can be thought of as drawing from an urn containing $N - L$ blue balls and L red balls. The probability of getting K red balls at the first K drawings can be calculated as follows:

$$p(topK) = \prod_{n=1}^{K} \frac{L+1-n}{N+1-n} = \frac{L!(N-K)!}{(L-K)!N!} \tag{4}$$

The last equation allows one to calculate the probability to have documents belonging to the correct category at the K top positions of the ranking with a random drawing. This can give a measure of the significance of the results obtained.
Two different Precision vs Recall curves have been obtained for noisy and clean version of the data respectively. The curves are close to each other and the Wilcoxon test shows that the difference is due to statistical fluctuations with a confidence level of 95%. In other words, the overall performance loss is negligible and the system appears to be robust with respect to the noise introduced in the data. For the sake of simplicity, the curves can be resumed in a single measure called average Precision (avgP), i.e. the average of π values along the curve. The higher the average Precision, the better the system. The comparison of the results obtained over clean and noisy versions of the same texts (see Table 2) shows that the average loss in terms of avgP when passing from clean to noisy texts is 8.7%.
Using equation (4), it is possible to calculate the probability of the results being obtained randomly: in all of five rankings related to the different categories (in the noisy version) the first 12 positions are occupied by documents belonging to the actual category under examination. This corresponds to a probability around 10^{-10} for each ranking and thus to a probability of $\sim 10^{-50}$ for the fact that all of them achieve such result. This suggests that the results obtained cannot be attributed to randomness or to the fact that the test set size is relatively small.
The performance loss changes depending on the category considered. Crude and grain seem to be more damaged by noise than the other categories, but this is due to the fact that they contain several

Category	avgP (clean)	avgP (noisy)	avg dist.
acquisitions	100.0%	99.0%	0.07
crude	99.5%	78.5%	0.64
earnings	100.0%	98.4%	0.29
grain	100.0%	84.2%	0.37
money-fx	95.7%	91.4%	0.20
average	99.0%	90.3%	0.31

Table 2: Average Precision and average distance from optimal transcription.

documents where the extraction process has a very low recognition rate (∼7% on average). In some cases, the recognition rate is 0% and the performance loss can thus be attributed to extraction process problems rather than to noise. This is confirmed by the fact that all other documents belonging to crude and grain (where the extraction process performance is average) occupy the top positions of the ranking. It is interesting to remark that even in the case of the earnings category, where the WER (TER) is 76.4% (49.4%), the categorization performance is close to the clean text case. The reason is that most errors are due to the misrecognitions of digits (very frequent in the earnings documents) and these have no effect on the categorization. On the other hand, the few words that are correctly recognized are sufficient for good categorization. It seems then to be important not only how many words are recognized, but also which words are recognized.
It is not easy to relate the extraction process performance and the categorization results. While the relation is evident for extreme cases (the 0% recognition rate of the above mentioned documents), it is no longer clear for intermediate situations. For the earnings category, the average recognition rate is low, but the categorization avgP is very high. Moreover, in several categories it is possible to find documents where the recognition rate is lower than the average, but still the transcription appears at the top ranking positions. A measure of the recognition performance better related to the categorization performance is thus necessary.

3.3 Information Gain Plan

The results presented in the previous subsection show that WER and TER are not meaningful performance measures in the context of noisy text categorization. The solution can be a measure that takes into account not only the percentage of words recognized, but also how much such words are representative of the category the document belongs to. A measure of such property is given by the *Information Gain* (IG) [2]:

$$IG(k) = H(c) - p(k)H(c|k) - p(\bar{k})H(c|\bar{k}) \tag{5}$$

where $IG(k)$ is the Information Gain of term k, $p(k)$ and $p(\bar{k})$ are the fractions of documents where term k is present and absent respectively and $H(.)$ is the entropy. The IG is high for terms appearing in several documents and being significantly related to few categories. After applying stopping and stemming to both original text and transcription, there are two sets: $I = \{t_1, t_2, \ldots, t_{|I|}\}$ contains the terms of the original text, $I^* = \{t^*_1, t^*_2, \ldots, t^*_{|I^*|}\}$ contains the terms of the transcription. This allows one to derive two measures:

$$\Phi(IG) = \frac{\sum_{i:t_i \in I \cap I^*} \min(tf(i), tf^*(i)) \cdot IG(i)}{\sum_{\{k:t_k \in I\}} tf(k) \cdot IG(k)} \tag{6}$$

$$\Psi(IG) = \frac{\sum_{i:t_i \in I \cap I^*} \min(tf(i), tf^*(i)) \cdot IG(i)}{\sum_{\{k:t_k \in I^*\}} tf^*(k) \cdot IG(k)} \tag{7}$$

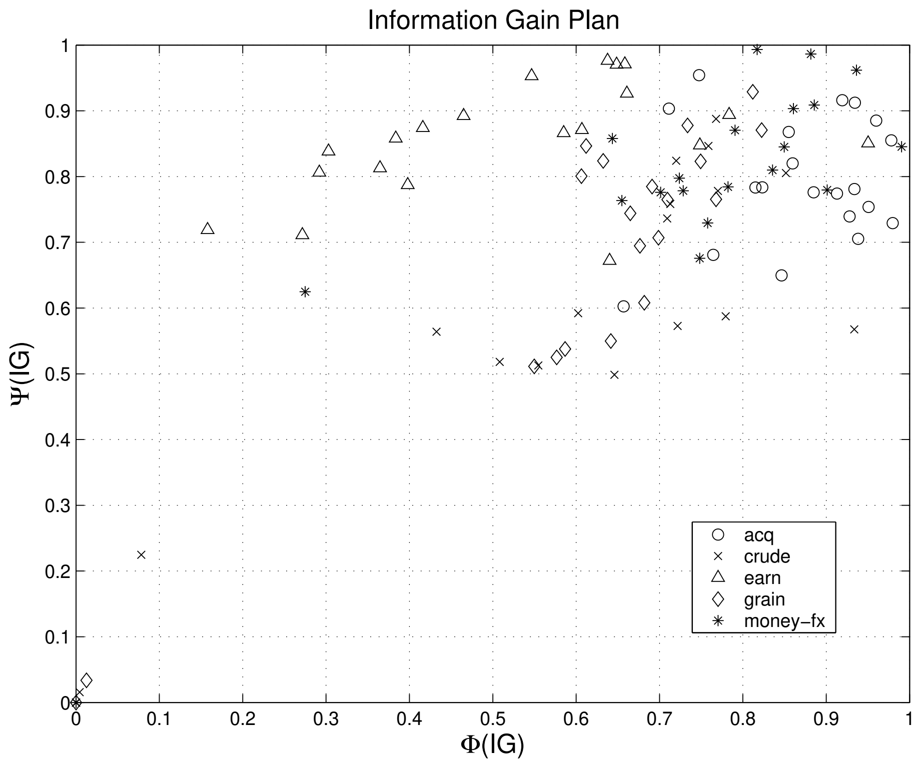

Figure 1: Information Gain Plan.

where $tf(i)$ and $tf^*(i)$ are the number of times term i appears in the clean and noisy text respectively. The measure $\Phi(IG)$ can be interpreted as the fraction of IG preserved through the recognition and of $\Psi(IG)$ as the fraction of transcription IG that is not the result of an error.
On a plan where Φ and Ψ are the coordinates, the point $(1,1)$ corresponds to an extraction process where all the information contained in the clean text has been preserved and no information other than the one contained in the original source has been added. On such a plan, the documents of classes for which the performance loss is lower are expected to be closer, on average, to the $(1,1)$ point. This is shown in figure 1 and the values of the average squared distance per class are collected in table 2. On average, the higher the average squared distance, the higher the performance loss when passing from clean to noisy version of the class texts. This seems to suggest that the IG based measure is better correlated with the categorization performance than the simple WER or TER. The reason is that the IG based measure accounts not only for the amount of words (terms) correctly recognized, but also for the amount of information useful to the categorization they represent.

4 Conclusions

This work presented a system for the categorization of noisy text. By noisy it is meant any text obtained (through an extraction process affected by errors) from sources different from digital text. Our results show that the categorization is robust with respect to an average WER of 50%.
In order to explain such results, an extraction process performance measure different from WER has been proposed. The measure is based on Information Gain and accounts for the amount of information useful to the categorization preserved through the extraction process.
As collections of several kinds of data (speech recordings, handwritten documents, slides, etc.) from

where texts can be extracted are more and more numerous, the categorization of noisy texts can be of great benefit. Several future works are possible. The categorization can be used as a step for further processing of the data the noisy text is extracted from (e.g. a category dependent language model can be selected to improve the recognition performance). Moreover, the possibility of processing in the same way texts extracted from different media can help to manage databases where different kinds of media appear.

Acknowledgements The author wishes to thank F.Camastra, D.Grangier, M.Keller, I.Lapidot and J.M.Odobez for their help. This work is supported by the Swiss National Science Foundation through the National Center of Competence in Research on Interactive Multimodal Information Management (IM2).

References

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval.* Addison Wesley, 1999.
[2] T. Joachims. *Learning to Classify Text using Support Vector Machines.* Kluwer, 2002.
[3] Selfcitation. Offline recognition of large vocabulary cursive handwritten text. *accepted for publication by IEEE Trans. PAMI*, 2004.