Sketching a “low-cost” text-classification technique for text topics in English

Pascual Cantos Gómez
Universidad de Murcia (Spain)
pcantos@um.es

Abstract

The aim of this paper is to sketch a potential methodology for automatic text classification which allows text topic discrimination as a prior step to new case assignment to previously established text topics. Such case assignment will be performed by means of Discriminant Function Analysis based on a series of easily-computable linguistic parameters, in order to reduce computational costs.

Keywords: automated text classification, discriminant function analysis, classification functions, English language, text topics.

Resumen

Esbozo de una técnica de clasificación de “bajo coste” según temáticas para textos en lengua inglesa

El objetivo de este artículo es esbozar una posible metodología para la clasificación automática de textos que permita la discriminación temática, como paso previo a la asignación de casos nuevos de textos a temáticas previamente establecidas. Dicha clasificación/asignación temática se implementa mediante el análisis discriminante y se sustenta en una serie de parámetros lingüísticos de fácil obtención, con el fin de reducir costes computacionales.

Palabras clave: clasificación automática de textos, análisis discriminante, funciones clasificatorias, lengua inglesa, categorías textuales.

1. Introduction

The widespread use of Internet has contributed to the existence of a great amount of information and documents which are within the user’s reach.

Internet traffic is approximately doubling each year. This growth rate applies not only to the entire Internet, but also to a large range of individual institutions (Coffman & Odlyzko, 2002).

The massive availability of documents on the web has triggered an increasing interest among the information retrieval community in looking for new methodologies for automatic text-knowledge extraction in order to organise and index texts in a more efficient way (Kenji, Ishiguro & Fukushima, 2001; Meadow et al., 2007). Automated text categorisation/classification into predefined categories has witnessed a booming interest, due to the increased availability of documents in digital form and the ensuing need to organize them (Sebastiani, 2002).

2. Automated text classification

In this section, we shall briefly survey the most common approaches and strategies to automatic text classification (ATC).

One of the earliest methods applied to ATC is Rocchio algorithm (Ittner, Lewis & Ahn, 1995; Cohen & Singer, 1999). This algorithm makes the "bag-of-words" assumption: it takes a document as a simple collection of words with varying frequency, ignoring the relative ordering of these words. More precisely, it treats a document as a vector where each entry is the term-frequency of that word in the text. Since such a vector would be extremely long and computationally very expensive, simple techniques are used to eliminate certain words from being used as features: words with extremely low occurrence along with functional words[1], as they are of little relevance to the categorization task. Rocchio algorithm computes a "prototype" vector for each class of text and uses the cosine distance of a document vector to the prototype of each class to determine the categorization, where the largest cosine distance implies the smallest angular distance and is chosen as the closest match. Joachims (1999) showed that the basic Rocchio algorithm is not particularly well suited to the task of categorization, but it serves as a good baseline and as a starting strategy of other similar approaches.

Joaquims (1999) himself proposed a more sophisticated strategy to ATC based on Rocchio algorithm assumption: Support Vector Machines (SVMs). SVMs are applied to the vector space model to find a category profile which produces the lowest probability error on document classification, achieving high accuracy ratings, about 86%, on the Reuters Corpus.[2] The main drawback of SVMs is its training time: quadratic in the numbers of training examples, to classify each document thousands of support vectors could be involved, and the time is usually too high.

Another extension of Rocchio algorithm was proposed by Schapire, Singer and Singhal (1998) and Lam and Ho (1998) and based on learning algorithms, produced good results on a number of standard test collections. However, both approaches relevantly increase the complexity of the basic model and increase time and computational cost.

A renewed interest in the Rocchio formula can be found in Moschitti's (2003) Profile-based text Classifiers (PBC). PBCs are characterized by a function based on a similarity measure between the synthetic representation of each class in the training set and the incoming document. Both representations are vectors, and similarity is traditionally estimated as the cosine angle between the two. The description of each target class is called profile, that is, a vector summarizing all training documents of that class. Vector components are called features and refer to independent dimensions in the similarity space. The PBC methodology increases accuracy and reduces the searching space of parameters, and suggests a simple and fast estimation procedure for deriving the optimal parameter (Moschitti, 2003).

Other less Rocchio-dependent approaches include k-Nearest Neighbor algorithm (kNN). Yang (1994) proposed this example-based text classifier that makes use of a document to document similarity estimation. It is also known as an instance-based classification method, and has been an effective approach to a broad range of pattern recognition and text classification problems (Dasarathy, 1991; Yang, 1999; Yang et al., 2000). However, the algorithm requires the calculation of all the scalar products between an incoming document and those available in the training set. Thus time complexity is rather high too.

Cohen and Singer (1999) also offer a non-Rocchio approach: a profile based one, RIPPER, using co-occurrences and multi-unit-words. Their algorithm decides to what extent the context (co-occurrences) of a word contribute actually to the target document classification. As it is based on profiles, it can be very fast, but it has a noticeable learning time. Moreover, given the complexity to derive the suitable multi-unit-words, it is not clear if it can be applied on a large scale.

Hierarchical text categorization (HTC) approaches have recently attracted a lot of interest, since they have been shown to bring about equal or similar classification accuracy while allowing significant time savings at both learning and classification time (Fagni & Sebastiani, 2007). HTC methods typically select negative examples: given a category *c*, its negative training examples are by default identified with the training examples that are negative for *c* and positive for the categories sibling to *c* in the hierarchy.

Succinctly, the central problem for all classification methodologies above is the high dimensionality of the feature space (the large number of features and vector space involved). This makes classification a complex and expensive process in time and computer space. Even worse, this leads also to sparse feature sets describing the documents: the number of features may grow so fast with the number of documents that in each new document too few features with a known predictive value reappear. Many features without predictive value might disturb the algorithm. On the opposite side, for sufficiently large numbers of documents the probability of reappearance of features goes up but, paradoxically, that large number might itself give problems too, as most texts will be classified as belonging to the most frequent topic(s), ignoring less frequent or rare topics. And this is erroneous too.

3. Research goal

Our aim is not to evaluate existing ATC methods, but trying to bring together existent ATC "philosophies" and come up with an inexpensive and easy-computable ATC strategy, even for non-Natural Language Processing experts.

We intend to use a classification technique based on feature strategies. However, and this is distinct to the feature-methodologies above, we do not aim at designing a complex learning algorithm that extracts features from text samples in order to create a feature vector space for each class of text or text-topic. Our starting approach is much simpler than that, we shall start with an *a priori* set of features; we shall take easy-computable linguistic features, that have already been discussed and/or used previously, with various degree of efficiency, to identify text typologies, topics and or genres, though not specifically to ATC (Yang & Pedersen, 1997; Baker & McCallum, 1998; McCallum et al., 1998; Nigam et al., 2000; Sebastiani, 2002; Guyon & Elisseeff, 2003; Debole & Sebastiani, 2005; among others). The research goal is to explore the accuracy of ATC into homogeneous topic-groups under chosen parameters and by means of standard statistical classification techniques.

As a secondary aim, we shall try to determine which of these easy-computable linguistic variables have a greater influence on determining the belonging of a particular text to a text-topic category and whether these variables work better in isolation or in combination. We shall try to find out which variables or groups of variables discriminate more reliably among text-topic categories.

4. Methodology

4.1. Corpus

We decided to build our own corpus for the present study following Stamatatos, Fakotakis and Kokkinakis (2000), Lee (2001), Lee and Swales (2006) and Axelsson (2010), who argue that the use of already existing corpora not built for text genre/topic detection – as in the case of the Brown Corpus or the British National Corpus (BNC)[3]– raise several problems since such categories may not be stylistically and lexically homogeneous. In addition, the way the corpus data will be used is both qualitative and quantitative (Lee, 2001). Consequently, our research corpus is an *ad hoc* corpus, consisting of five text-topic categories (see Table 1), equivalent to Lee's (2001: 49) notion of "genres":

> We can see that the categories to which texts have been assigned in existing corpora are sometimes genres, sometimes subgenres, sometimes 'super-genres' and sometimes something else. This is undoubtedly why the catch-all term 'text category' is used in the official documentation for the LOB4 and ICE-GB5 corpora. Most of these 'text categories' are equivalent to what I am calling 'genres'.

This is not an exhaustive research, but a small-scale study. The categories of the current corpus are just a small number of the total text-topic categories or genres (even subgenres) which may be identified in English. Samples were taken from open access academic journals and scientific magazines.

We chose five text topics, namely "Ecology", "Music", "Oncology", "Physics" and "Religion". For each category, ten written texts were collected from different websites. The corpus was compiled following these four criteria:

1. Real-world text in electronic form.
2. Raw text; neither linguistic tags nor any other manual/machine text-processing restrictions are set.
3. Texts were cleaned up removing HTML, javascripts, links, navigational information, advertisement, etc.
4. Text length was limited to 1,500 words.

In order to achieve this 1,500-word requirement, original texts, often over 5,000 words, had to be shortened. Thus, links, authors' details, acknowledgements, bibliography and references were left out, as well as abstracts, introductions and conclusions since these were considered very general sections offering little discriminating ground among text categories. As for the rest of the text, whole pages from each section were randomly selected in order to cater for representativeness. Overall ten texts were collected for each text category (see Table 1).

Text category	Tokens per text	Nº of texts	Total tokens
Ecology	1,500	10	15,000
Music	1,500	10	15,000
Oncology	1,500	10	15,000
Physics	1,500	10	15,000
Religion	1,500	10	15,000
Total		50	75,000

Table 1. Corpus composition.

4.2. Variables

Different to previous feature-based methodologies, our approach takes an *a priori* set of features. The variables or features used in the present study have been selected under the criteria that these are linguistic, quantitative and require minimal computational costs, which for our purpose means that no tagged or parsed text is required. Thus, syntactic and structural features have been kept out of the current study.

A thorough literature review on text/topic/genre categorization/classification (Yang & Pedersen, 1997; Baker & McCallum, 1998; Nigam et al., 2000; Sebastiani, 2002; Guyon & Elisseeff, 2003; to name a few) suggest a great variety of linguistic variables and parameters for text classification. However, these variables can be broadly summarised and categorised under three main headings as discussed below: (i) punctuation variables, (ii) lexical distribution variables, and (iii) most frequent words.

4.2.1. Punctuation variables

These variables belong to the so-called "token-level measures" (Honoré, 1979; Stamatos, Fakotakis & Kokkinakis, 2001). That is, the input text is considered as a sequence of tokens grouped in sentences. This level is based on the output of the sentence boundary detector. Token-level measures have been widely used in both text genre detection and authorship attribution research since they can be easily detected and computed (Stamatatos, Fakotakis & Kokkinakis, 2000 & 2001; Putnins et al., 2005; Pinto, Jiménez-Salazar & Rosso, 2006). As Stamatatos, Fakotakis and Kokkinakis (2000) show, there are cases where the frequency of occurrence of a certain punctuation mark could be used alone for predicting a certain text genre. For example, an interview is usually characterized by an uncommonly high frequency of question marks. Similarly, Quirk et al. (1985) examine punctuation marks and conclude that these are beyond the level of the word and up to the level of the sentence.

In this research, we have considered eight measures of punctuation:

1. Periods.
2. Commas.
3. Semicolons.
4. Colons.
5. Dashes.
6. Pairs of parentheses.
7. Exclamation marks.
8. Question marks.

4.2.2. Lexical distribution variables

Lexical distribution measures are surface level characteristics as well as several linguistic phenomena of the vocabularies of the corpora in order to identify important variations of language use among them (Verspoor, Bretonnel Cohen & Hunter, 2009). The lexical distribution measures variables we shall be looking at are: "sentence length", "vocabulary richness" and "readability indexes".

4.2.2.1. Sentence length

Following Holmes (1994), Gómez Guinovart and Pérez Guerra (2000) establish two measures for sentence length as a measure of extensive use in stylometric works of authorship attribution. These are: (i) words per sentence; and (ii) characters per sentence.

4.2.2.2. Vocabulary richness

Various measures have been proposed for capturing the richness of the vocabulary of a text. Biber (1988) uses two of the most common measures: "type-token ratio" and "word length". In addition, we shall also include: "hapax legomena" and "hapax dislegomena" in the terms explained below. The vocabulary richness measures examined in this study are:

1. Type-token ratio. This ratio measures the number of different lexical items in a text, as a percentage. In the present study, the measure used is that proposed by Scott (1999): the "standardised type/token ratio" (hereon STTR). STTR is computed every *n* words as wordlist goes through each text file. By default, *n* = 500. In other words, the ratio is calculated for the first 500 running words, then calculated afresh for the next 500, and so on to the end of the text or corpus. A running average is computed, which means that you get an average type-token ratio based on consecutive 500-word chunks of text.
2. Word length. This is the mean length of words in orthographic letters. Longer words are said to convey in general more specific and specialised meanings than shorter ones and Zipf (1949) showed that words become shorter as they are more frequently used and more general in meaning.
3. Another measure will be used along with word length in orthographic characters following Karlgren and Cutting (1994): "long word count" – that is, words with more than six characters.
4. Hapax Legomena. These are once occurring words and indicators of style. They are related to vocabulary richness and precision.

5. Hapax Dislegomena. These are words which occur only twice in a text (Holmes, 1994).

4.2.2.3. Readability indexes

We decided to use two measures of readability grades, following previous studies that suggest that these measures may be used as powerful differentiators between text types (Karlgren & Cutting, 1994; Gómez Guinovart & Pérez Guerra, 2000):

> Automated Readability Index = 4.71 * letters_per_word + 0.5 * words_per_sentence – 21.43
>
> Coleman-Liau Index = 5.89 * letters_per_word – 0 .3 * sentences_per_100_words –15.8

4.2.3. Most frequent words

Variables	Features		
Punctuation variables	1. Periods		5. Hyphens
	2. Commas		6. Parentheses
	3. Semicolons		7. Exclamations
	4. Colons		8. Questions
Lexical distribution variables	9. Words / Sentence		14. Hapax Legomena
	10. Characters / Sentence		15. Hapax Dislegomena
	11. Standt.TTR		16. Automated Readability Index
	12. Word Length		17. Coleman-Liau Index
	13. Long Word Count		
Frequency of occurrence of the 30 most frequent words	18. The	28. With	38. Are
	19. Of	29. He	39. Not
	20. And	30. Be	40. His
	21. A	31. On	41. This
	22. In	32. I	42. From
	23. To	33. That	43. But
	24. Is	34. By	44. Had
	25. Was	35. At	45. Which
	26. It	36. You	46. She
	27. For	37. 's	47. They

Table 2. Variables analysed.

We included the frequency of occurrence of the 30 most common words from the BNC. This is a variable proposed in many studies, particularly of authorship attribution, using a set of typically 30 or 50 high frequency words (Burrows, 2002; Hoover, 2004; Stein & Argamon, 2006; Argamon, 2008, etc.). Stamatatos, Fakotakis and Kokkinakis (2000) found the best performance (error rate = 2.5) at the 30 most frequent words of the BNC corpus, comprising the following words[6]: “the”, “of”, “and”, “a”, “in”, “to”, “is”, “was”, “it”, “for”, “with”, “he”, “be”, “on”, “I”, “that”, “by”, “at”, “you”, “’s”, “are”, “not”, “his”, “this”, “from”, “but”, “had”, “which”, “she”, “they”. Finally, Table 2 summarises the total number of variables used (47 overall):

4.3. Statistics

We need to assign individual texts, for which several variables have been measured, to certain groups or text topic categories that have already been identified in the corpus. The statistic technique we shall use for this aim is discriminant function analysis (DFA, hereafter; Cantos Gómez, 2013: 104-112).

DFA involves the prediction of a categorical dependent variable (text topic category) by one or more independent variables (47 variables defined above). It uses the set of independent variables to separate cases based on groups one defines; the grouping variable is the dependent variable and it is categorical (text-topic; namely, Ecology, Music, Oncology, Physics and Religion). DFA creates new variables based on linear combinations of the independent set that one provides. These new variables are defined so that they separate the groups as far apart as possible. How well the model performs is usually reported in terms of classification efficiency, that is, how many texts would be correctly assigned to their groups using the new variables from DFA. The new variables can also be used to classify a new set of cases. If DFA is effective for one set of data, the classification table of correct and incorrect estimates will yield a high percentage of correct ones.

DFA shares all the usual assumptions of correlation, requiring linear and homoscedastic[7] relationships, and interval or continuous data. It also assumes the dependent variable is categorical. It is broken into a two-step process:

1. Significance testing: a test is used to check whether the discriminant model as a whole is significant; and,
2. Classification: if the test reveals significance, then the individual independent variables are assessed to see which differ significantly in mean by group and these are used to classify the dependent variable.

Once the corpus was compiled, the next step was to compute the texts by means of a standard concordance program[8] to obtain the values for each of the variables which respect to each individual text.

4.4. Data analysis

According to our research goal, our aim is twofold:

1. To predict the categorical dependent variable ("domain/text topic"), to *a priori* defined groups, for 47 independent variables, and to check if the discriminant model as a whole is significant; and
2. If the model shows significance, then to assess the individual independent variables, selecting those variables with a greater discriminant capacity and to generate a predictive discriminant model to classify new cases.

With more than one independent variable, it is very time consuming to do all the calculations manually, so we shall present the results of the DFA using SPSS[9], commenting only on those data tables which are relevant to our analysis.

First, we obtain some preliminary descriptive data (means and standard deviation scores on each variable for genres (Ecology, Music, Oncology, Physics and Religion), and the overall mean standard deviations on each variable, which is not relevant commenting here. Next, a tolerance test is undertaken to assess the viability of all independent variables prior to analysis. SPSS produces eight variables which fail the tolerance test (Table 3). Consequently, these variables are excluded as predictors in the DFA.

	Intra-group variance	Tolerance	Minimum tolerance
Automated Readability Index	4.078	.000	.000
at	23.756	.073	.000
's	19.840	.084	.000
this	4.078	.157	.000
but	1.747	.209	.000
had	5.367	.076	.000
they	2.533	.170	.000

Table 3. Variables that do not pass the tolerance test.

The tolerance is an indication of the percentage of variance in the predictor that cannot be accounted for by the other predictors; hence, very small values indicate that a predictor contains redundant information. The minimum required tolerance level is 0.001.

The next relevant table (see Table 4) gives information on the ratio of importance of the dimensions (functions) which classify cases of the dependent variable. There is one "eigenvalue" for each discriminant function. For two-group DAF, there is one discriminant function and one eigenvalue, which account for 100% of the explained variance; for three-group DAF there will be two discriminant functions and two eigenvalues, etc. Note that the number of discriminant functions is equal to the number of groups we want to classify minus one. If there is more than one discriminant function, the first will be the largest and most important one, the second the next most important in explanatory power, and so on.

Function	Eigenvalue	% of variance	Cumulative %	Canonical correlation
1	20.075	49.2	49.2	.976
2	12.562	30.8	80.0	.962
3	5.059	12.4	92.4	.914
4	3.087	7.6	100.0	.869

Table 4. Eigenvalues.

The "canonical correlation" is a measure of the association between the groups formed by the dependent and the given discriminant function. When the canonical correlation is zero, there is no relation between the groups and the function. However, when the canonical correlation is large, then there is a high correlation between the discriminant functions and the groups; that is, it tells you how much each function is useful in determining group differences.

The data relative to our model reveals that the first function explains 49.2% of the variance, whereas the second one does 30.8%, the third 12.4% and the forth only 7.6%. Consequently, most of the discriminating power for the model is associated with the first three discriminant functions. The canonical correlation indexes show a high correlation between the discriminant functions and the groups (0.976, 0.962, 0.914 and 0.869). That is, each function contributes significantly to determining group differences.

Now we shall look at the significance testing, in order to know whether our discriminant model as a whole is significant or not. SPSS performs the

"Wilks' lambda test" (see Table 5). This multivariate test is a statistic that tells us about the "fit" of the data. The first two functions show high significant *p*-values (*<0.05*), in contrast functions 3 and 4 are less positive; however, we can say that the model is a good fit for the data; that is, the predicting variables[10] used discriminate positively.

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 4	.000	239.344	156	.000
2 through 4	.003	157.045	114	.005
3 through 4	.040	239.344	74	.149
4	.245	38.009	36	.378

Table 5. Wilks' Lambda test.

Table 6 gives the classification table that we get by selecting that option in the SPSS dialog box. It gives information about actual group membership vs. predicted group membership. The overall percentage correctly classified equals 80 – that is, 80% of original grouped cases were correctly classified by means of the four discriminant functions inferred from the data provided. This speaks very much in favour of our model and its predictive power.

	Genre	Predicted Group Membership					
		Ecology	Music	Oncology	Physics	Religion	Total
Original Count	Ecology	8	1	0	0	1	10
	Music	0	8	1	0	1	10
	Oncology	1	0	9	0	0	10
	Physics	0	1	0	9	0	10
	Religion	1	2	1	0	6	10
%	Ecology	80,0	10,0	,0	,0	10,0	100,0
	Music	,0	80,0	10,0	,0	10,0	100,0
	Oncology	10,0	,0	90,0	,0	,0	100,0
	Physics	,0	10,0	,0	90,0	,0	100,0
	Religion	10,0	20,0	10,0	,0	60,0	100,0

Table 6. Classification results.

To further explore the robustness of our model, we performed a cross-validation. Cross-validation is a standard tool in analytics and is an important feature for helping to develop and fine-tune models. It is used after creating a model, in order to ascertain its validity. It assesses how the results of a statistical analysis will generalize to an independent set of data. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. Cross-validation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset ("training set"), and validating the analysis on the other subset ("testing set"). Table 7 gives the classification table with cross-validation and the overall percentage of correctly classified text-topic samples is still very promising: 70%, particularly if we consider the computational and temporal low cost. Major problems are found in correctly assigning Ecology texts. These are mainly grouped under Ecology or Religion.

	Genre	Predicted Group Membership					
		Ecology	Music	Oncology	Physics	Religion	Total
Original Count	Ecology	4	1	1	0	4	10
	Music	0	8	1	0	1	10
	Oncology	1	1	8	0	0	10
	Physics	0	1	0	9	0	10
	Religion	1	2	1	0	6	10
%	Ecology	40,0	10,0	10,0	,0	40,0	100,0
	Music	,0	80,0	10,0	,0	10,0	100,0
	Oncology	10,0	10,0	80,0	,0	,0	100,0
	Physics	,0	10,0	,0	90,0	,0	100,0
	Religion	10,0	20,0	10,0	,0	60,0	100,0

Table 7. Cross-validation classification results.

5. Modelling text-classification

Once the DFA has turned out to be positive, we can try a stepwise procedure instead. This will allow us to assess each individual independent variable in order to select the best predictor or set of predictors. SPSS now optimises and simplifies the model and outputs a new model with similar predictive power, however, using as few predictors as possible (see Table 8), namely: number of semicolons; standardized type-token ratio (STTR); Coleman Liao Index (CLI); and occurrences of "in".

Predictors	Function			
	1	2	3	4
Semicolons	.242	1.076	-.083	.181
STTR	.760	-.033	.671	-.032
CLI	.876	-.051	-.709	.251
Token_in	-.658	-.537	.512	.780

Table 7. Best predictors and coefficients.

Curiously, if we compare this simplified model with the one above using all independent variables, the CLI variable passes the tolerance test, whenever it combines with the other three variables (semicolons, STTR and "in"), becoming a positive classification model.

A visual representation (see Figure 1) of the model shows how the five groups (genres) separate out from one another using these four predictors just. The group centroid stands for the prototypal sample of each text topic, and the individual text-topic samples gather around its centroid. The model measures for each text-topic sample the distances between the sample and the different centroids; the least distance between a sample and a centroid determines its membership.

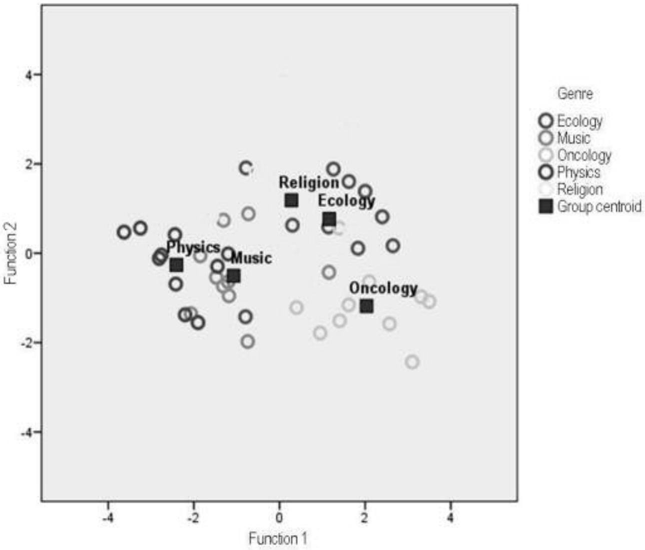

Figure 1. Domain distribution based on functions 1 and 2.

A further usability of DFA is that, once we have selected the variables with a greater discriminant capacity, it is possible to generate a predictive discriminant model to classify new cases. By means of selecting SPSS option "Fisher Function Coefficients", we are given a table (Table 8) with a constant value and a number of coefficients for each of the best predictors (semicolons, STTR, CLI and "in") with reference to each linguistic domain (Ecology, Music, Oncology, Physics and Religion.

			Genre		
	Ecology	Music	Oncology	Physics	Religion
Semicolons	2.021	.977	1.009	.995	2.053
STTR	8.868	8.655	9.509	7.784	9.130
CLI	4.010	2.391	4.054	2.281	2.895
Token_in	-.222	.281	-.064	.271	-.015
(Constant)	-214.174	-201.849	-243.088	-165.416	-222.756

Table 8. Discriminant function coefficients.

This gives five equations, one for each genre:

Ecology	= -214.174+(2.021*Semicolons)+(8.868*STTR)+(4.010*CLI)+(-0.222*In)
Music	= -201.849+(0.977*Semicolons)+(8.855*STTR)+(2.391*CLI)+(0.281*In)
Oncology	= -243.088+(1.009*Semicolons)+(9.509*STTR)+(4.054*CLI)+(-0.064*In)
Physics	= -165.416+(0.995*Semicolons)+(7.784*STTR)+(2.281*CLI)+(-0.271*In)
Religion	= -222.756+(2.053*Semicolons)+(9.130*STTR)+(2.895*CLI)+(-0.015*In)

To illustrate the applicability of these equations, we can take, for example, a randomly chosen 1,500-token text out of Ecology, Music, Oncology Physics and/or Religion journals/magazines. Imagine that after computing it, we get the following values for the variables semicolons, STTR, CLI and "in":

Number of semicolons = 3

Standardized type-token ratio = 43.21

Coleman Liao Index = 6.12

Occurrences of "in" = 40

Using the discriminant equations above and instantiating the values for semicolons, STTR, CLI and "in", we can calculate the scores of the three discriminant functions:

Ecology	= -214.174+(2.021*3)+(8.868*43.21)+(4.010*6.12)+(-0.222*40) = 190.736
Music	= -201.849+(0.977*3)+(8.855*43.21)+(2.391*6.12)+(0.281*40) = 200.937
Oncology	= -243.088+(1.009*3)+(9.509*43.21)+(4.054*6.12)+(-0.064*40) = 193.073
Physics	= -165.416+(0.995*3)+(7.784*43.21)+(2.281*6.12)+(-0.271*40) = 198.715
Religion	= -222.756+(2.053*3)+(9.130*43.21)+(2.895*6.12)+(-0.015*40) = 195.0277

The randomly chosen text with: semicolons = 3; STTR = 43.21; CLI = 6.12; and "in" = 40, will be assigned to the genre, related to one of the five equations above, that has the largest resulting value. So maximising the five coefficients, we find that this text is most likely to be a music text, as Music is the highest resulting coefficient (200.937); and in second place, it would be classified under Physics (198.715). Similarly, the least likely group membership would be Ecology (190.736), as the coefficient obtained in the ecology equation is the lowest one.

6. Some final remarks

This research has tried to identify a set of linguistic markers that discriminates effectively among the text-categories proposed for this study, under the assumption that stylistic differentiation of texts will enable automatic text classification. That is, if text samples can, at a more than chance rate, be differentiated from each other, it will mean that the set of variables accurately catch stylistic variation. Therefore, on facing a new text sample, and been provided with the linguistic data of the linguistic parameters, the model may correctly assign the "unknown" text sample to the text category it belongs to.

As testing ground we have used our purpose-built corpus which comprised five text categories and a total of 15,000 words per category. Results have shown that the set of linguistic variables proposed, in addition to being easily identified and computed, can accurately discriminate among text categories. The DFA offered a 70% accurate classification[11] of text samples into the categories analysed. In addition, the model favours stylistic differentiation. It is interesting to note that these categories (Ecology, Music, Oncology, Physics and Religion), are often considered as a whole and comprised within the broader genre of "academic prose" suggesting some degree of homogeneity.

A further contribution of DFA is the possibility of creating new models for classifying new cases, using not all predicting variables, but just a limited number of them: the best predictors, with similar or identical discriminatory power. This reduces the dimensionality of the data and produces "low cost" models, not only computationally speaking, but also with regard to the time and effort spent on the data collection process.

As for the set of variables, more work needs to be done on reducing the number of parameters. With the aim of economizing, while keeping accuracy, in future research we will work on a model able to efficiently discriminate among a greater number of text categories based on an even simpler set of parameters.

Of course, two of the major shortcomings of this approach are: (a) its systemic circularity. That is, the model relies on a list of linguistic features resulting from the stylistic(al) analysis of genres; next, the model checks the presence/absence of such features in the individual texts, so the theoretical basis of the framework is quite circular; and (b) the limited size of the corpus – therefore, its representativeness. Future research should therefore focus on a greater and more representative corpus catering for as many text genres as possible to check its generalisation power.

[Paper received 25 February 2013]
[Revised paper accepted 22 July 2013]

References

Argamon, S. (2008). "Interpreting Burrows's delta: Geometric and probabilistic foundations." *Literary and Linguistic Computing* 23: 131-147.

Axelsson S. (2010). "The normalized compression distance as a file fragment classifier". *Digital Investigation: The International Journal of Digital Forensics & Incident Response* 7: 524-531.

Baker, L.D. & A.K. McCallum (1998) "Distributional clustering of words for text classification" in *Proceedings of SIGIR'98, 21st ACM International Conference on Research and Development in Information Retrieval*, 96-103. New York: ACM Press.

Biber, D. (1988). *Linguistic Features: Algorithms and Functions in Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Burrows. J. (2002). "Delta: A measure of stylistic difference and a guide to likely authorship". *Literary and Linguistic Computing* 17: 267-287.

Cantos Gómez, P. (2013). *Statistical Methods in Language and Linguistic Research.* Sheffield: Equinox.

Coffman, K.G. & A.M. Odlyzko (2002). "Growth of the Internet" in I.P. Kaminow & T. Li (eds.), *Optical Fiber Telecommunications IV B: Systems and Impairments*, 47-93. Norwell, MA: Kluwer Academic Publishers.

Cohen. W.W. & Y. Singer. (1999). "Context-sensitive learning methods for text categorization". *ACM Transactions on Information Systems* 17: 141-173.

Dasarathy, B. (1991). *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*. Los Alamitos, CA: IEEE Computer Society Press.

Debole, F. & F. Sebastiani (2005). "An analysis of the relative hardness of reiters-21578 subsets". *Journal of the American Society of Information Science and Technology* 56: 584-596.

Fagni, T. & F. Sebastiani (2007). "On the selection of negative examples for hierarchical text categorization" in *Proceedings of the 3rd Language Technology Conference*, 24-28. Pozlan: PL.

Gómez Guinovart, X. & J. Pérez Guerra (2000). "A multidimensional corpus-based analysis of English spoken and written-to-be-spoken discourse". *Cuadernos de Filología Inglesa* 9: 39-70.

Guyon, I. & A. Elisseeff (2003). "An introduction to variable and feature selection". *Journal of Machine Learning Research* 3: 1157-1182.

Holmes. D.I. (1994). "Authorship attribution". *Computers and the Humanities* 28: 87-106.

Honoré, T. (1979). "Some simple measures of richness of vocabulary". *ALLC Bulletin* 7: 172-177.

Hoover, D.L. (2004). "Testing Burrows's delta". *Literary and Linguistic Computing* 19: 453-475.

Ittner, D.J., D.D. Lewis & D.D. Ahn (1995). "Text categorization of low quality images" in O. Maimon & L. Rokach (eds). *Proceedings of SDAIR-95*, 301-315. Las Vegas: ISRI University of Nevada.

Joachims, T. (1999). "Advances in Kernel methods – Support vector learning" in B. Schölkopf, C.

Burges & A. Smola (eds.) *Making large-Scale SVM Learning Practical*, 41-56. Mass: MIT Press.

Karlgren. J. & D. Cutting (1994). "Recognizing text genres with simple metrics using discriminant analysis". URL: http://eprints.sics.se/56/01/cmplglixcol.pdf [10/12/12]

Kenji, T., Y. Ishiguro & T. Fukushima (2001). "Opinion information retrieval from the Internet". *IPSJ NL* 144: 75-82.

Lam, W. & C.Y. Ho (1998). "Using a generalized instance set for automatic text categorization" in Y. Chiaramella (ed.) *Proceedings of the 21st ACM SIGIR Conference*, 81-89. New York: ACM.

Lee. D. (2001). "Genres, registers, text-types, domains and styles: Clarifying the concept and navigating a path through the BNC jungle". *Language Learning and Technology* 5: 37-72.

Lee, D. & J. Swales (2006). "A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora". *English for Specific Purposes* 25: 56-75.

McCallum, A., R. Rosenfeld, T. Mitchell & A. Ng (1998). "Improving text classification by shrinkage in a hierarchy of classes" in *Proceedings of the Fifteenth International Conference on Machine Learning (ICML'98)*, 412-420. New York: ACM Press.

Meadow, C.T., B.R. Boyce, D.H. Kraft & C.L. Barry (2007). *Text Information Retrieval Systems.* San Diego, CA: Academic Press.

Moschitti, A. (2003). "A study on optimal parameter tuning for Rocchio text classifier" in F. Sebastiani (ed.) *Advances in Information Retrieval, 25th European Conference on IR Research*, 420-433. Heidelberg: Springer Verlag.

Nigam, K., A. McCallum, S. Thrun & T. Mitchell (2000). "Text classification from labeled and unlabeled documents using EM". *Machine Learning* 39: 103-134.

Pinto, D., H. Jiménez-Salazar & P. Rosso (2006). "Clustering abstracts of scientific texts using the transition point technique" in A.F. Gelbukh (ed.) *CICLing 2006. LNCS*. 536-546. Heidelberg: Springer.

Putnins, T.J., D.J. Signoriello, S. Jain, M.J. Berryman & D. Abbott (2005). "Who wrote the letter to the Hebrews?" in A. Bender (ed.), *Proceedings of SPIE: Complex Systems 6039*, 1-13. Brisbane: University of Brisbane Press.

Quirk. R., S. Greenbaum, G. Leech & J. Svartvik (1985). *A Comprehensive Grammar of the English Language*. Longman: London.

Schapire, R.E., Y. Singer & A. Singhal (1998). "Boosting and Rocchio applied to text filtering" in W. Bruce Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson & J. Zobel (eds.), *Proceedings of SIGIR-98*, 215-223. New York: ACM Press.

Scott, M. (1999). *WordSmith Tools 3.0.* Oxford: Oxford University Press

Sebastiani, F. (2002). "Machine learning in automated text categorization". *ACM Computing Surveys* 34: 1-47.

Stamatatos, E., N. Fakotakis & G. Kokkinakis (2000). "Text genre detection using common word frequencies". URL: http://www.cs.mu.oz.au/acl/C/C00/C00-2117.pdf [03/09/12]

Stamatatos, E., N. Fakotakis & G. Kokkinakis (2001). "Computer-based authorship attribution without lexical measures". *Computers and the Humanities* 35:193-214.

Stein, S. & S. Argamon (2006). "A mathematical explanation of Burrows's delta". *Proceedings of Digital Humanities.* URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.71.8771&rep=rep1&type=pdf [08/05/11]

Verspoor, K., K.B. Bretonnel Cohen & L. Hunter (2009). "The textual characteristics of traditional and open access scientific journals are similar". *BMC Bioinformatics* 10: 1-9.

Yang, Y. (1994). "Expert network: Effective and efficient learning from human decisions in text categorization and retrieval" in W.B. Croft & C.J. van Rijsbergen (eds.), *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 13-22. New York: Springer Verlag.

Yang, Y. (1999). "An evaluation of statistical approaches to text categorization". *Information Retrieval* 1: 67-88.

Yang, Y., T. Ault, T. Pierce & C. Lattimer (2000). "Improving text categorization methods for event tracking" in E. Yannakoudakis, N.J. Belkin, M.K. Leong & P. Ingwersen (eds.), *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval*, 55-72. New York: Springer Verlag.

Yang, Y. & J. Pedersen (1997). "A comparative study on feature selection in text categorization" in *Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97)*, 412-420. New York: ACM Press.

Zipf, G.K. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge: Addison-Wesley.

Pascual Cantos Gómez is Professor in the Department of English Language and Literature at the University of Murcia, Spain. His main research interests are Corpus Linguistics, Quantitative Linguistics, Computational Lexicography and Computer Assisted Language Learning. He is a member of the LACELL (*Lingüística Aplicada a la Computación, Enseñanza de Lenguas y Lexicografía*) Research Group at the University of Murcia.

NOTES

[1] Also known as stop words. These are excluded from some language processing tasks, usually because they are viewed as non-informative or potentially misleading. Usually they are non-content words like conjunctions, determiners, prepositions, etc.

[2] The most commonly used corpora for ATC is the Reuters-21578 Corpus. Assembled in 1987, the corpus contains 21,578 documents that appeared in the Reuters news media during that year. URL: http://www.daviddlewis.com/ resources/testcollections/reuters21578

[3] British National Corpus (BNC), URL: http://www.natcorp.ox.ac.uk

[4] Lancaster-Oslo/Bergen Corpus, URL: http://khnt.hit.uib.no/icame/manuals/lob/ INDEX.HTM

[5] British component of the International Corpus of English, URL: http://www.ucl.ac.uk/english-usage/projects/ice-gb

[6] This list was taken from a non-lemmatised list of the most frequent words of the BNC – see URL: http://www.comp.lancs.ac.uk/ucrel/bncfreq/flists.html

[7] Those that have equal statistical variances.

[8] A computer program that lets you create word lists and search natural language text files for words, phrases, and patterns.

[9] Statistical Package for the Social Sciences (SPSS), URL: http://www-01.ibm.com/ software/es/analytics/spss

[10] Except for those variables that did not pass the tolerance test (number of long words and average word length) and were excluded from the model.

[11] Cross-validation classification results.