

OCR - Classification Interaction Mathematical Model

Jean-Marie Brodin

A2iA, 40 bis rue Fabert 75007 Paris France
jbr@a2ia.com

Abstract

The topic of this article belongs to the field of semantic classification of documents generated by an OCR process. We propose models of the composite system made of a semantic classifier and an OCR process, with 3 different types of relations between these two processes, one with a deterministic OCR output, and two with probabilistic OCR outputs. We show the connections between these 3 OCR categories. We propose an explanation of the origin of the robustness of the global system with respect to the OCR errors, based on an analogy with communication theory. The models can be used to estimate a priori the performances of the global system, knowing the intrinsic difficulty of the semantic classification problem, and the OCR error rate. It can also be used to compare the efficiencies of the different ways to connect the OCR with the semantic classifier.

Keywords: Semantic classifier, OCR-generated document

1. Introduction

The interest in the Information Retrieval field, has kept on increasing, mainly because of the ever growing number of numeric documents available on Internet. However, besides this huge database of digitalized documents, there also coexists a non negligible source of informations, namely, the set of paper, machine printed or handwritten, documents. The techniques of semantic or content-based document classification are generally based on the statistical search for keywords whose presence or absence are characteristic of a given semantic class. A whole category of basic classification algorithms are thus based on the joint probabilities $P(C_i, w_j)$ or $P(C_i, \bar{w}_j)$, of the presence or absence of word w_j in document class C_i .

We consider here the problem of semantic document classifiers, taking as input characteristics, the set of words generated by an OCR process applied on a graphical document. The processing of the paper documents requires their transformation to a digital representation, through the use of an OCR process. Because the output of the OCR

process is entailed with errors, and also, differs in nature from the natural representation of a document (seen as a set of words belonging to a lexicon), directly generated from an edition software, there is a need to optimize the relation between the OCR output, and the semantic classification process. The paper cited in annex (Information Retrieval for OCR Documents : A Content-based Probabilistic Correction Model), is an instance of the active research in this field. Like this article, as far as the author knows, most articles propose essentially algorithms to improve the synergy between the lower level OCR process, and the upper level document classification, or Information Retrieval process.

This document differs in the sense that it searches some ways to estimate the performance of a composite system made of a low-level recognition process, and an upper-level classification process, from the performances of the parts and the intrinsic difficulty of the document classification problem, independently of specific OCR or classification algorithms.

This paper is therefore the first in a serial that plans to address the following problems :

- provide a tool allowing a priori estimation, or even boundary-setting to the performance of a document semantic classification problem for paper documents, given the intrinsic difficulty of the classification problem, and the OCR performances.

- provide a theoretical model upon which algorithms, as well as experimental reports could be compared, and gain some insight in the global process, as well as in the interconnection process between the OCR and the classifier processes.

- For deterministic OCRs, that we will call "type I", providing only a single candidate word, with no confidence score, the theoretical model can be obtained from communication theory, and doesn't require any modeling of the OCR process itself. We will base this model on the analogy between on one hand a noisy communication channel, and on the other hand, an OCR generated document, the OCR process being seen as a source of noise, damaging an ideal document. The ideal document being here a digital document generated by an edit software.

- For probabilistic OCRs, that we will call respectively "type II" and "type III", providing a single word with its recognition score, or a full lattice of candidates with their scores, the optimal model that naturally stands out is the Bayesian model. Type II mathematical model may be seen as just a restriction of type III.

- For deterministic (Type I) OCRs, the mechanism that makes the whole system robust towards OCR errors is from this analogy easily understandable : it is the same principle as error correcting codes in communication theory. We would like to have the same insight for probabilistic OCRs (Types II-III). In the limit of a perfect OCR, Type II-III models should tend towards type I. However, this is not the case. We show that a connection between the two types of models remains possible, but that it requires the introduction of a parameter ϵ in the type II model : that is, probabilities should be made close to 0 or 1, but not exactly equal to these two values, in order to have type II OCR model tend towards type I mathematical model in the limit of a perfect OCR. This way, we gain some insight also in the global process in the case of probabilistic OCRs : its robustness also (at least partially) comes from the same principle as error correcting codes. Type III OCRs prove to be more robust than types I or II, especially for low performance OCRs, because of another mechanism, that we also analyze.

For type I OCRs, we will be able to provide a mathematical model that will not require any other a priori knowledge about the OCR process, than its error rate. But, for type II and III, we will need to make assumptions about the OCR generation process, and we will propose to model it with the help of a Gaussian noise model, and a Bayesian classifier. This assumption will allow us to build a full mathematical model linking the three quantities defined above. However, because the solution requires the computation of integrals of the error function, and of its inverse, that have no analytical form, we will provide only a numerical solution of the "Monte-Carlo" type, unlike in type I, where we don't even need to provide a model of the recognizer.

We will use these numerical solutions, to construct curves showing the great trends of the models. All the curves show the relation between the OCR recognition rate, and the document classification rate. The curves essentially differ with respect to the intrinsic difficulty of the (ideal) semantic classification problem, which is characterized by the number of significant keywords (lexicon size), and with respect to the OCR type. We call here "ideal semantic classification problem", the case where the documents to be classified, either are digital documents, generated, for instance, by a text editor, or are generated from a scanned paper document, by a perfect, errorless OCR process.

The observed trends relate :

- 1) The performance of the document classifier, with respect to the OCR error rate and the OCR type, for a given difficulty of the ideal problem.
- 2) The performance of the document classifier, with respect to the OCR error rate and the difficulty of the ideal problem, for a given OCR type
- 3) The influence of a parameter called "epsilon" on the performances of the type II OCR.

2 Mathematical description of the three types of OCR :

2.1 Type-I OCR global mathematical model :

We derive here the great steps of the computation of the error rate, η^C , of the classifier, as a function of the OCR error rate, η , the number of keywords, n (also called the lexicon size), and the number of classes, M :

- An "ideal" document is a set of keywords, which is a subset of the lexicon. In a space with N dimensions, in which each dimension corresponds to a word of the lexicon, each ideal document corresponds to a vertex of an hypercube of side 1, and dimension N (For the sake of simplicity, we don't consider here multiple instances of the same word, because it does'nt contain any fundamental difference). Example :

$$D_j = (O_1, \dots, O_n, O_{n+1}^-, \dots, O_N^-) \quad (1)$$

means that the document D_j contains the (ideal) keywords (O_1, \dots, O_n) , and not keywords (O_{n+1}, \dots, O_N) . This document D_j will thus be represented by a vertex of the hypercube, having for coordinates, 1 for the n first dimensions, and 0 for the last $N - n$ dimensions. The characteristic space for documents is essentially made of the vertices of the N -dimensional hypercube. This topology is the one that corresponds to the ideal classification problem. Let's note here that the vertices represent the actual document in the recognition phase for type I OCRs. However, the space of the documents representing the classes, even in the type I case, is not limited to the vertices of the hypercube. Indeed, the vector representing the class is made of the conditional probabilities $P(C_j|O_k)$, and these values, even in the type I case, are not restricted to 0 or 1, but can take any real value inside this interval. However, it doesn't change here the essence of the reasoning, to consider that the class vector corresponds to a vertex. We will call here the "ideal class vertex", the vertex of the hypercube closest to the document class characteristic vector. Edges link vertices that differ only with respect to the presence or absence of one keyword : if E_j is the edge linking the two vertices :

$$V_k = \{\dots, O_j, \dots\} \quad (2)$$

and

$$V_{k'} = \{\dots, \bar{O}_j, \dots\} \quad (3)$$

$$d(V_k, V_{k'}) = 1 \quad (4)$$

The natural distance between two characteristic vectors V_i and V_j is Hamming's distance.

- If the OCR process was not entailed with any errors, then, each physical document word would be correctly recognized, and the list of O_i generated by the OCR would match exactly the list corresponding to the document class. Therefore, the document generated by the OCR could be located, in the hypercube, at the vertex corresponding to the document class.
- Each OCR error on a word, moves the point representing the physical document, away from the ideal class vector, by one vertex, and corresponds to one given edge. If the OCR makes d errors, then the physical document vertex will be at a hamming distance d away from the true vector. If the error rate of the OCR is η , and if the dimension of space (the number of keywords) is n , then, the probability of generating a document located at a Hamming distance d away from the ideal class vertex, is :

$$P(d) = \eta^d \cdot (1 - \eta)^{(n-d)} \quad (5)$$

- A natural document classification algorithm consists in assigning to the physical document, the ideal class having the smallest Hamming distance with that physical document. For such an algorithm, each ideal class is assigned with a domain, like an "attraction basin", which may be approximated by a sphere with radius r . The error rate is therefore the percentage of cases where the physical document corresponding to a given class C_j , lies outside the hypersphere of the class C_j . This rate is given by the formula :

$$\eta^C = \sum_{d>r} C_d^n \cdot \eta^d (1 - \eta)^{(n-d)} \quad (6)$$

where C_d^n is the number of vertices located at distance d , from the class vector.

- One can see that the above formula requires the estimation of the average radius, r , of the hyperspheres of the classes. This radius is a function of the dimension of space, n , and of the number of classes. The "volume", or the number of vertices, of an hypersphere of radius r , in a Hamming space of dimension n , is :

$$G(n, r) = \sum_{k=0}^r C_k^n \quad (7)$$

the total volume of the Hamming space, being equal to 2^n

If M is the number of classes, we can consider that the equation :

$$M \cdot G(n, r) = 2^n \quad (8)$$

gives implicitly r as a function of M and n .

- The estimation of η^C , as a function of η , n and M , therefore, goes as follows :
 - 1) Compute r as a function of M and n , such that : $M \cdot G(n, r) = 2^n$.
 - 2) Compute the sum over $n \geq d \geq r$, of : $C_d^n \eta^d (1 - \eta)^{(n-d)}$. This sum is the error rate η^C of the document classifier :

$$\eta^C = \eta^C(n, r, \eta) = \sum_{n \geq d \geq r} C_d^n \eta^d (1 - \eta)^{(n-d)} \quad (9)$$

and :

$$\eta^C = \eta^C(n, M, \eta) \quad (10)$$

since :

$$r = r(n, M) \quad (11)$$

Plots of η^C , as a function of η , for increasing values of n , or increasing values of M , show clearly how the keywords play the role of error correcting codes : these curves are basically made of two linear parts : the function starts flatly, until it rapidly turns down, to decrease linearly : that is, the classifier can absorb the errors of the OCR, up to a certain extent : its characteristics are nearly independent from the OCR performances, until the OCR error rate reaches a certain threshold, above which the classifier error rate is directly and linearly related to the OCR error rate.

2.2 Type-III OCR global mathematical model

In this paragraph, we want to relate the quantity :

$$P(C_k | X_1, \dots, X_i, \dots, X_n) \quad (12)$$

to :

$$P(O_k | C_j) \quad (13)$$

and to :

$$P(O_k | X_i) \quad (14)$$

The $P(C_k|X_1, \dots, X_i, \dots, X_n)$ are the probabilities that a document made of the vector of patterns $X_1, \dots, X_i, \dots, X_n$ belongs to class C_k .

The $P(O_k|C_j)$ are the conditional probabilities to find the word "k", in document class "j". These probabilities are the "ideal" probabilities, that would be the ones found for digital documents, that is, documents not generated by an OCR process, but rather directly, for instance, by a text editor.

Finally, the $P(O_k|X_i)$ are the "OCR" probabilities that the object with form, or characteristic vector X_i belongs to the class O_k .

Therefore, we want to relate our document classification scores, on one hand to the purely semantic characteristics of the problem, described by the $P(O_k|C_j)$, and on the other hand, to the OCR characteristics of the problem, described by the $P(O_k|X_i)$.

As a first step, we will construct the conditional probabilities to find the shape X_i , on document class C_j , as a sum over all possible paths :

$$P(X_i|C_j) = \sum_k P(X_i|O_k)P(O_k|C_j) \quad (15)$$

Where the C_j are the document classes, the O_k are the objects making up the document, and the X_i are the forms produced by the generating process, from the O_k . In a second step, using Bayes inversion formula in the case of n stochastic variables :

$$P(C_j|X_1, \dots, X_n) = \frac{\Pi_i P(X_i|C_j) P^n(C_j)}{P(X_1, \dots, X_n)} \quad (16)$$

and replacing in this equation, $P(X_i|C_j)$ by its value given in the former equation, we get :

$$\Rightarrow P(C_j|X_1, \dots, X_n) = \Pi_i \left(\sum_k P(O_k|X_i) \frac{P(X_i)}{P(C_j)} P(O_k|C_j) \right) \frac{P^n(C_j)}{P(X_1, \dots, X_n)} \quad (17)$$

Where the $P(O_k|X_i)$ is the probability vector produced by the OCR process applied on form X_i .

$$\Rightarrow P(C_j|X_1, \dots, X_n) = \frac{P^n(C_j) \Pi_i P(X_i)}{P(X_1, \dots, X_n) P^n(C_j)} \Pi_i \left(\sum_k P(O_k|X_i) P(O_k|C_j) \right) \quad (18)$$

If the X_i may be considered as stochastically independent, then :

$$P(X_1, \dots, X_n) = \Pi_i P(X_i) \quad (19)$$

$$\Rightarrow P(C_j|X_1, \dots, X_n) = \Pi_i \left(\sum_k P(O_k|X_i) P(O_k|C_j) \right) \quad (20)$$

which is the relation we were looking for.

This equation merges together two conditional probabilities with different origins : the first set of probabilities, the $P(O_{k(i)}|X_i)$ are generated by the low level OCR process, while the second set, the $P(O_{k(i)}|C_j)$ corresponds to the upper level document classification process.

We will now use this relation to link the performances of the classifier with the performances of the OCR, and to the ideal problem difficulty.

However, we will first start by a discussion around this equation.

First, we recognize in the set $P(O_k|X_i)$, a vector belonging to the n dimensional space we met earlier, in the Type I OCR. However, here, the vector can span the whole space. It is not limited to the vertices of the hypercube.

2.3 Type-II OCR global mathematical model

Type II OCRs may be considered as a special case of Type III OCRs, for which a single element is selected in the sums $\sum_k P(O_k|X_i)P(O_k|C_j)$

$$\Rightarrow P(C_j|X_1, \dots, X_n) = \Pi_i P(O_{k(i)}|X_i) P(O_{k(i)}|C_j) \quad (21)$$

We call "naive choice rule", the rule that generates a type II output from a type III output, by selecting, for each graphical word X_i , the class O_j with the highest score p_{ij} .

As we said earlier, when the OCR process is embedded in a higher level classification process, the naive choice rule is not necessarily the optimal. This is due to the presence of the second term, $\Pi_i P(O_{k(i)}|C_j)$. Because of this term, the combination with maximum $\Pi_i P(O_{k(i)}|X_i)$ is not necessarily the one that maximizes $\Pi_i P(O_{k(i)}|X_i) P(O_{k(i)}|C_j)$. Consider for instance the cases where the $P(O_{k(i)}|C_j)$ are equal only to 0 or 1. Then, the combination that maximizes $\Pi_i P(O_{k(i)}|X_i)$ may well be a combination for which $\Pi_i P(O_{k(i)}|C_j) = 0$, that is, for that combination, the global score is 0. Another way to see this consists in taking the problem backwards, and in considering the only combinations for which $\Pi_i P(O_{k(i)}|C_j) = 1$, which are the combinations corresponding to the ideal documents. For each of these combinations, there is a combination of O_j that maximizes the OCR score $\Pi_i P(O_{k(i)}|X_i)$, and therefore, the optimal document D_m is the one among the allowed combinations, that maximizes this OCR score. This is obviously not necessarily the one that would have been selected by the naive choice rule. In the case of low performance OCR,

the probability is high that for each word X_i , the class with highest score is not the correct one. But the probability is low, since all the OCR processes are independent, that these randomly generated top-score classes correspond to one of the ideal classes. Therefore, in these cases, using the naive rule, automatically leads to a classification error.

3 Comparison between the 3 types of OCRs :

- type II process is equivalent to type I, only if the $P(O_k|C_j)$ are made equal to $\epsilon > 0$ or $1 - \epsilon$ instead 0 or 1. This " ϵ " value corresponds to an artificial dispersion of the conditional probabilities $P(C_j|O_k)$, which is indeed obtained if the OCR process is used during the learning phase that generates the lexicon.

- The simulations show indeed that the type III OCR is more efficient than type II, especially when the performance rate of the OCR is low.

Type III differs from type II essentially through the fact that the naive choice rule tends to discard correct solutions, while type III keeps them. Type III diverges from type II when the OCR error rate increases : as we will see, this trend is also confirmed by the simulations.

4 Discussion of the simulation curves given in annex :

- Simulations 1 and 2 show the influence of the number of keywords. As long as the OCR error rate lies below a given threshold, the errors of the OCR are fully absorbed by the classifier.
- Simulations 2 and 3, show the equivalence of OCRs I and II, in case the value of ϵ is small but non-zero (equal to 0.1), and the number of keywords is large (here, 8).
- Simulations 1 and 5, show the equivalence, for a small number of keywords (here, 4), of type I OCR, and type II when ϵ is equal to 0. Simulation 6 show that type II OCRs are indeed more robust than type I, when the difficulty of the ideal problem increases.
- Simulations 3 and 7 show the influence of the ϵ factor for the type II OCR.
- Simulations 6 and 8 show that the type III OCR is more robust than type II and type I, and show its largest ability to absorb the errors of the OCR.
- Simulations 3 and 9, show the same trend as above, with 8 keywords instead of 4. Therefore, the type III OCR shows a clear superiority above the two other types of OCRs, for a large range of OCR recognition rate, and ideal problem difficulty.
- Finally, simulations 2 and 4 seem to show that type I OCRs performances stop increasing when the difficulty of the ideal problem gets below a given threshold.

5 Conclusion :

Using the analogy between noisy communication channels, and noisy documents generated by an OCR process, we propose to see the mechanism that makes a semantic document classifier operator robust towards the noise introduced by the OCR process, as analogous to the mechanism of error correcting codes in communication theory. We studied this property, and compared the performance of the global system, for three types of OCRs, distinct by the nature of their outputs, deterministic or probabilistic, and in this last case, providing a single candidate class with its confidence score, or a full lattice of probabilities. We showed the relation between the deterministic and the probabilistic models, and proposed that the error correcting codes mechanism is in both cases at the root of the robustness of the global system towards the errors introduced by the OCR. We have then shown from the simulations that the probabilistic model that outputs a full lattice of probabilities is more robust and provided reasons for this property. Finally, we provide the results of a mathematical model that can compute the performance level of the document classifier, for a given difficulty of the ideal problem, and a given value of the OCR error rate.

Références

- [1] Rong Jin ; Chengxiang Zhai ; Hauptmann Alex G., "Information Retrieval for OCR Documents : A Content-based Probabilistic Correction Model", *Electronic Imaging Conference (EI'03), Document Recognition and Retrieval Conference DRR(X)*, Santa Clara, CA, January 20-24, 2003, vol. 5010, pp. 128-135.
- [2] J. Bass, *Eléments de Calcul Probabilités (théorique et appliqué)*. Masson et Cie.
- [3] J. Oswald, *Théorie de l'Information ou Analyse Diacritique des Systèmes*. Masson et Cie.

6 Annex : Curves generated by the numeric simulation

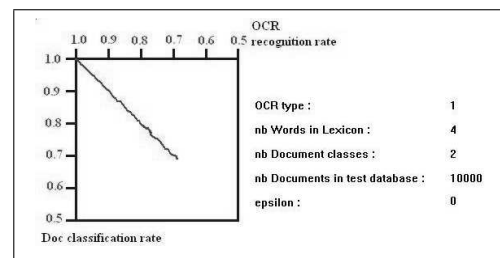


Figure 1. Simulation 1

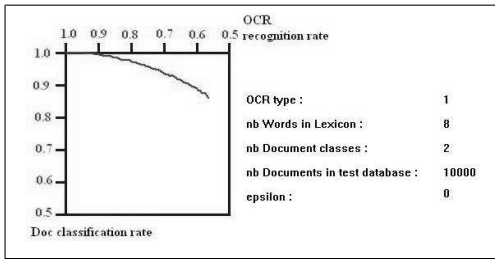


Figure 2. Simulation 2

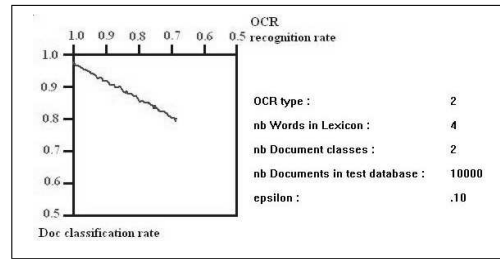


Figure 6. Simulation 6

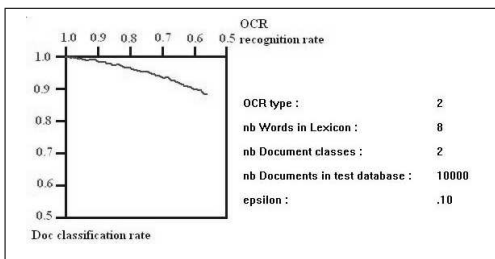


Figure 3. Simulation 3

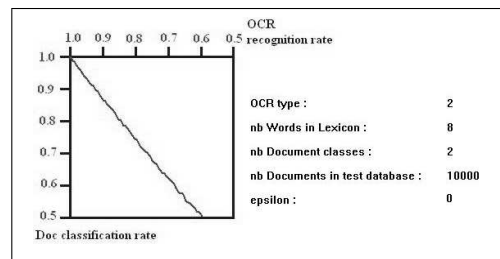


Figure 7. Simulation 7

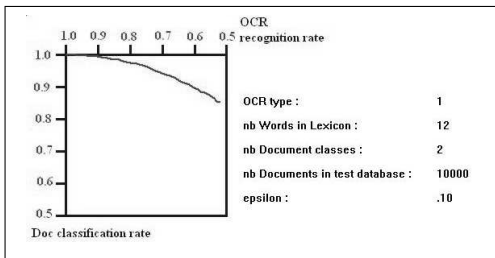


Figure 4. Simulation 4

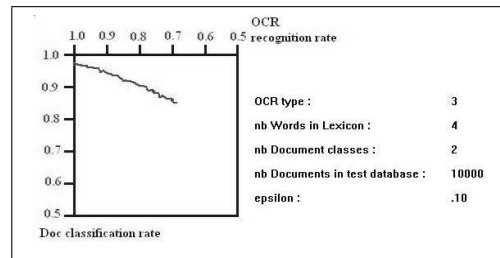


Figure 8. Simulation 8

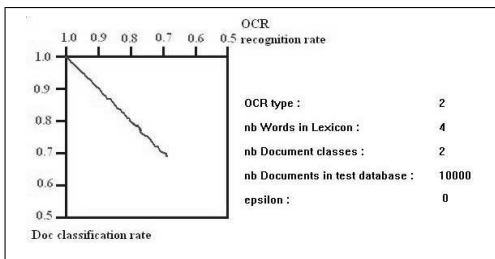


Figure 5. Simulation 5

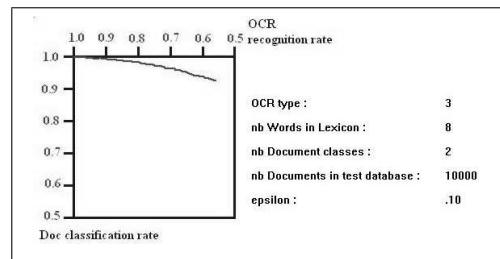


Figure 9. Simulation 9