The Influence of Order on a Large Bag of Words

Charles B. Prado
Research and Development Dept. — Globo Television Network, Rio de Janeiro
Email: charles.prado@gmail.com

Felipe M. G. França, Ramon Diacovo
Department of Systems Engineering and Computation — COPPE
Universidade Federal do Rio de Janeiro
Email: [felipe,ramon]@cos.ufrj.br

Priscila M. V. Lima
Department of Electronic Engineering and Computing
Universidade Federal do Rio de Janeiro
Email: pmvlima@gmail.com

Abstract

Text classification has been mostly performed through implicit semantic correlation techniques, such as latent semantic analysis. This approach however, has proved insufficient for situations where there are short texts to be classified into one or more from many classes. That is the case of the classification of statements of purpose of Brazilian companies, according to the around one thousand and eight hundred categories of the government administration detailment of National Classification of Economical Activities (CNAE), CNAE-Subclasses. The impact of the order of words in a text is evaluated by comparing the performance of three classifiers based on the weightless artificial neural model, WISARD. Results point to the need of combining semantic with syntactic information in order to improve the classifiers performance.

1 Introduction

Digital Economy, a term coined by Don Tapscott [9], requires adaptability from a business in order to survive in a very dynamic and competitive environment. This adaptability and agility are also required from the public administration in order to keep up with the pace of private businesses. Among other tasks, businesses need to be classified according to their economical activities. In Brazil, this classification is self-attributed in most of the cases. But self-classification may not be desirable, specially in the case of small businesses which do not have the resources to arrange private consultancy on the subject. Ideally, this classification should be performed by a group of expert human classifiers. However, if we consider business mutability in the Digital Economy, it is possible that human classification of economical activities becomes a bottleneck for the economical sector in the near future. Therefore, it is necessary to device a mechanism that automatically attributes economical activities to businesses from businesses data. The nature of such data may vary, but one should first consider using that already provided to the process carried out by humans. Even in this case, there are various sources of data: statements of purpose, trial balances, receipts, etc.

There are several international and national classification tables for economical activities [5], among them, there is the Brazilian table known as CNAE [1]. The CNAE-Subclasses table [2], that details the CNAE, is organized as a five-layer hierarchy. The layer of Sections is the uppermost of CNAE, followed by Divisions, Groups, Classes and, finally, Subclasses. Table 1 summarizes the structure of CNAE-Subclasses[1]. As an illustration, consider the subclass Culture of Rice, code 0111-3/01. It belongs to Section A (Agriculture, Cattle Raising, Forest Production, Fishing and Aquaculture), Division 01 (Agriculture, Cattle Raising and Related Services), Group 011 (Production of Seasonal Cultures) and Class 0111-3 (Culture of Cereals).

Aiming to research the viability of performing automated classification of businesses activities, the System for Automated Classification of Economical Activities (SCAE)

[1] From the second level onwards, the model is aggregative, i.e., the code of each grouping level incorporates that of the predecessor in the classification tree.

Table 1. Hierarchical Classification of CNAE 2.0

Name/Level	Groupings	Identifier
Section/1st	21	1-digit Alphabetical code
Division/2nd	87	2-digits Numerical code
Group/3rd	285	3-digits Numerical code
Class/4th	673	4-digits Numerical code + VD (verifier-digit)
Subclass/5th	1,301	7-digits Numerical code (incluing VD)

[3] intends to propose, compare and, if necessary, combines several classifiers based on Computational Intelligence techniques. Initially, the classifiers have as input the statement of purpose of businesses and output the code of one or more subclasses that correspond to the economical activities described in the document. It is worth noticing that these documents are not long, so the classifiers must be able to match it to a category (or categories) from very few words. As both the range of possible categories, 1301 subclasses, and the universe of words are very large, that task becomes extremely difficult. This work focuses on the study of classifiers based on the WISARD weightless neural network model [6]. In special we intend to investigate the gain from minimally adding information about the order of words. Section 2 outlines the main features of Weightless Artificial Neural Networks (WNNs). Section 3 describes the characteristics of the database of business statements of purpose used, as well as the three variations of the WISARD model investigated together with classification results. Section 4 concludes the article and points to future steps.

2 The WISARD Perceptron

Differently from most artificial neural network models, in which synaptic weights play the most important role in their functionalities, the WISARD perceptron is based on mimicking how axons and dendrites of biological neurons are interconnected. Although a McCullogh-Pitts' neuron deals with binary inputs, it produces an [0,1] output based on a weighted (i.e., synaptic strengths) sum of such inputs modulated by a non-linear threshold function. In contrast, each neuron in a WISARD perceptron, takes the functionality of a random access memory (RAM) and, being able to store 2^n bits, receives and produces binary values as well (see Fig. 1). This way, each neuron is able to learn and recognize n-bit words (*tuples*).

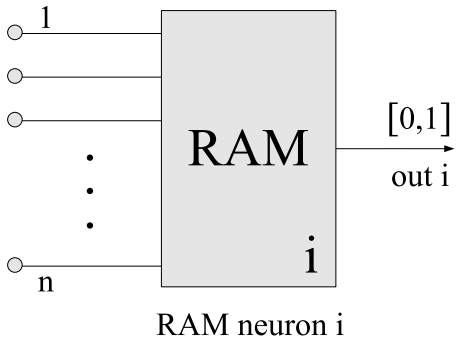

Figure 1. A RAM neuron

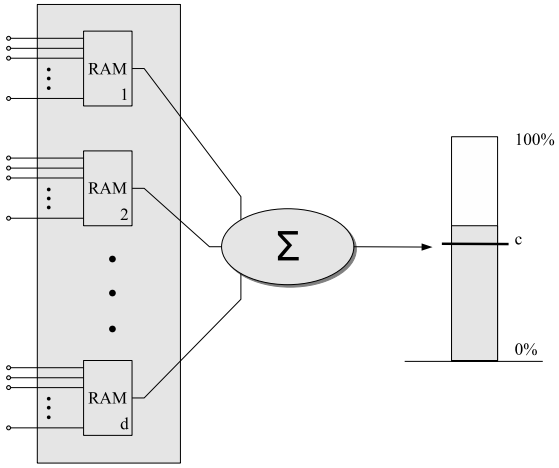

Figure 2. A WISARD discriminator

Training in both neural models is performed in the following ways: while a McCullogh-Pitts' neuron i learns by changes on its input weights w_{*i}, a RAM neuron i, having all its positions initialized with 0's, learns by writing '1' in the position addressed by the input tuple. Although RAM neurons are fast to train and recall, one can argue that the RAM neuron alone lacks generalization power since only previously learned tuples can be recognized. In order to overcome this drawback, RAM neurons are organized in a structure, called *discriminator*, where each neuron, of a total of d neurons, is responsible for the learning/recognition of a subset of a $n \times d$ input pattern. A discriminator is able to recognize a possibly unseen input pattern X through performing a simple summation of all its neurons' output bits. This graded response, submitted to a *confidence* level C, as shown in Fig. 2, could associate X to an already trained *class*.

The WISARD perceptron consists on an array of m, $m \in$

N^*, discriminators, each representing a different class of patterns. As illustrated by Fig. 3, an input pattern, after submitted to a shuffling function S, may be fed to:

(i) to one of the discriminators, during training phase, and;

(ii) to all m discriminators, during recognition phase, when all m responses are analyzed. Upon presenting a target pattern as input, in the case of single label classification, at most one class could be selected. Multiple label classification can be naturally performed upon training the target input pattern into the discriminators associated to their corresponding classes.

A recent successful application of the WISARD perceptron can be found in [7].

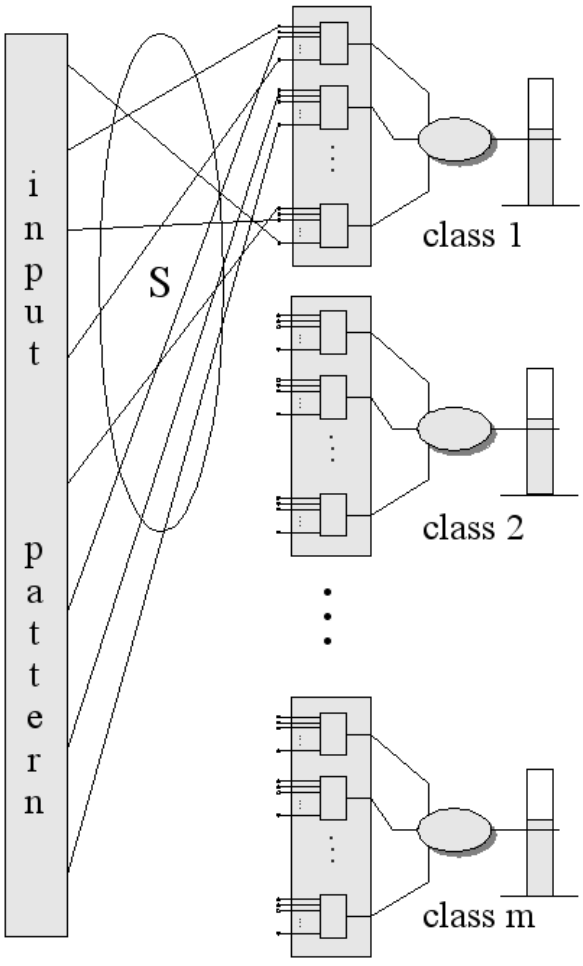

Figure 3. A WISARD perceptron

3 The WISARD Classifiers of SCAE

For our experiments, we used a database which consisted of 3264 statements of purpose of Brazilian companies, as well as 764 records from the Subclass (5^{th}) level of the CNAE table. On the training phase, the classifiers were fed with records from the CNAE table. The statements of purpose were then used to test the classifiers.

Initially, a pre-processing stage was applied to produce training and testing data sets. Firstly, only words appearing in the CNAE-Subclasses entries were considered. Thus, by using stop-words filtering and stemming processes, 1001 terms were obtained. Therefore, the resulting training set consisted of a matrix of dimensions 764x1001, while the testing set matrix has dimensions 3264x1001.

In both matrices, a position in the terms vector holds value '1' if the corresponding word is present in the text, otherwise its value is '0'. For the training set, the one containing the denominations of the Subclasses, an average of 8 words composed each vector. In many cases only 2 words composed the vector. On the other hand, the vectors of the testing set, that contained data from the statements of purpose, had as many as 70 words. This discrepancy may be explained by the fact that some businesses perform more than one economical activity.

3.1 Single WISARD (SiW)

The first classifier consists of a single WISARD (SiW). At the pre-processing stage, each term vector of 1001 elements was replaced by a 32x32 image, the last 23 remaining positions were filled with '0's. At the training stage, the images obtained for each of the 764 CNAE-Subclasses codes were assigned to a different WISARD discriminator. During the testing stage, all the 3264 statements of purpose were used. Likewise, each terms vector was converted into a 32x32 image. Each discriminator outputs the number of recognized words minus the number of missed words. The winning discriminator(s) is(are) the one(s) with the highest output. As a result, the SiW correctly classified 61% of the total of 3264 samples. For this analysis we considered only one of the winners (single-label classification).

3.2 Hits-only WISARD (HoW)

The Hits-only WISARD (HoW) modifies the SiW classifier by considering as output only the number of hits, that is, the HoW does not take into account the number of missed words. As a result, the HoW correctly classified 64% of the total of 3264 samples, thus representing an improvement of 3% over SiW result.

3.3 Ordered WISARD (OrW)

In both of the classifiers presented in subsections 3.1(SiW) and 3.2(HoW), the order of words appearing in the text was not taken into consideration. However, texts in natural language convey a lot of information in their structure. The least amount of information on structure that could be considered is the order of a word in the text. With the aim of evaluating the impact of such information on the performance of one of our classifiers, we modified the observation scheme of the testing stage of HoW. For each word presented to HoW in the testing (or recognition) stage, the list of winning discriminators was observed. It should be expected that, at each step, the number of winning discriminators would decrease until convergence. However, this would only occur in the case of single-activity businesses. So, the testing stage was modified to chose the winner before divergence started. This simple modification yielded an improvement of around 7% over HoW performance. That classifier will be addressed from now on as OrW. Figure 4 shows an example of text categorization using OrW. The first sentence of a given statement of purpose, in Portuguese, is *"Comércio varejista de máquinas, aparelhos e equipamentos eletrônicos, de uso doméstico e pessoal, equipamento de informática"* (Wholesale of machinery, electronic equipment and appliances, for domestic and personal use, informatics equipment). This sentence is followed by the word *"Manutenção"*, belonging to the next sentence inside the target text. The winning discriminator was defined when the word *"Manutenção"*, at position k, was presented to OrW. That word induced a divergence (from 1 to 40 positive discriminator answers).

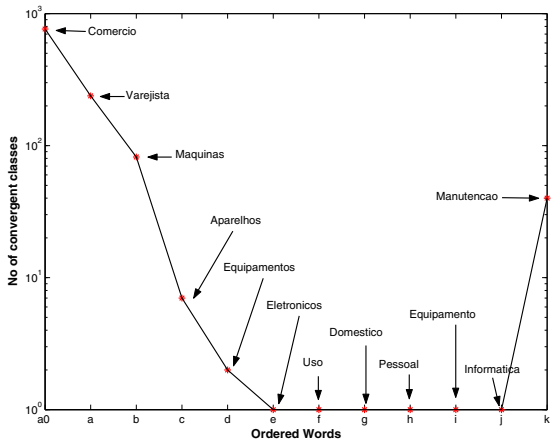

Figure 4. Example of convergence using OrW.

Table 2 summarizes the performance of the three classifiers.

Table 2. Performance by Classifier

Classifier	Results
SiW	61.3%
HoW	64.5%
OrW	71.8%

4 Conclusion and Future Work

We have presented and compared three classifiers based on the WISARD WNNs model for the task of attributing Subclasses codes to texts containing businesses statement of purposes. In order to improve the classification, the experiments suggest that the order of the words should be taken into consideration, given the small amount of words to indicate the categories in question. The same mechanism can be used to separate the pieces of text related to one economical activity.

As immediate future steps, we intend to include more complex syntactic information, starting from grammatical classes, attributed by a tagger [8, 4] and to apply the mechanism of OrW to decomposing text by activities. We also intend to further improve the classifier by including explicit knowledge about the domain of the different economical sectors, those that correspond to the Sections of CNAE.

Acknowledgements

The authors would like to thank *Receita Federal do Brasil* for funding this research through the SCAE project.

References

[1] Concla — Comissão Nacional de Classificações http://www.ibge.gov.br/concla/ (in portuguese). 2008.

[2] Concla — Comissão Nacional de Classificações http://www.ibge.gov.br/concla/revisao2007.php (in portuguese). 2008.

[3] LCAD — UFES — SCAE http://www.lcad.inf.ufes.br/index.php?option=com_content&task=view&id=17&itemid=40 (in portuguese). 2008.

[4] Treetagger http://www.ims.uni-stuttgart.de/projekte/corplex/treetagger/. 2008.

[5] United nations statistics division http://unstats.un.org/unsd/cr/ctryreg/. 2008.

[6] I. Aleksander, W. Thomas, and P. Bowden. Wisard, a radical step forward in image recognition. *Sensor Ver.*, 4:120–124, 1984.

[7] C. B. do Prado, F. M. G. França, E. Costa, and L. Vasconcelos. A new intelligent systems approach to 3d animation in television. In *CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 117–119, New York, NY, USA, 2007. ACM.

[8] P. Gamallo, A. Agustini, and G. P. Lopes. Clustering syntactic positions with similar semantic requirements. *Comput. Linguist.*, 31(1):107–146, 2005.
[9] D. Tapscott. *The Digital Economy*. Mcgraw-Hill, 1996.