STRESS ASSIGNMENT IN SPANISH PROPER NAMES

R. San-Segundo, J.M. Montero, R. Córdoba, J. Gutiérrez-Arriola

Grupo de Tecnologí a del Habla. Departamento de Ingenierí a Electrónica. UPM.
E.T.S.I. Telecomunicación. Ciudad Universitaria s/n, 28040 Madrid, Spain
lapiz@die.upm.es
http://www-gth.die.upm.es

ABSTRACT

In this paper, we propose an approach for Stress Assignment in Spanish Proper Names, based on a Multi-Layer Perceptron (MLP). When assigning stress to a word, we first analyse each vowel in the word and then calculate a *Stress-Confidence Measure* for it, using a MLP. The system will assign the stress to the vowel with the highest stress-confidence measure. In this paper we present and analyse different alternatives for the inputs to the Multi-Layer Perceptron. In all cases, we consider the number of vowels in the name and the vowel position in the word (taking into account only the vowels in the analysed word). For the rest of inputs, we consider a window of letters. These letters are obtained from the context of the vowel considered and from the word ending, in a similar way to [1]. We propose a *Discrimination Measure* to analyse the discrimination power for the different input configurations and we validate this measure and present the results obtained in each case. For the best configuration we obtain a 94.9% proper names correctly stressed (5.1% error rate). These results are compared to similar experiments using a *Memory based learning* approach (k-Nearest Neighbours).

Keywords: Stress Assignment, Spanish Proper Names, Multi-Layer Perceptron, k-Nearest Neighbour.

1. INTRODUCTION

This paper is concerned with a well-defined instance of linguistic pattern matching problems: the assignment of stress. In this work, we want to investigate if a Multi-Layer Perceptron is powerful enough to abstract the regularities governing the stress assignment. In Spanish, when a word is correctly stressed, there is no ambiguity for stress assignment. With a reduced set of simple rules it is possible to assign the stress for any Spanish word without any uncertainty. The problem grows up when we consider big databases with millions of proper names, e.g. banks and PTTs databases, and these companies want to provide automatic services using speech synthesis. These databases contain stress assignment errors that make the synthesiser do a wrong stress assignment. These errors were provoked by the fact that old databases did not accept words with accent mark in capital letters. These kind of services often use the user name during the interaction to provide a more friendly service. For example, they can make the greeting to be configurable, or use the name for user confirmation before giving his/her personal information (the account balance, or the number and type of telephone calls made during the last month). In these situations, a wrong stress assignment in the user proper name can produce two different effects. In the best case, the user identifies his/her name but he/she gets a bad impression from the system because he/she perceives an impersonal and uncertain service (the system can not pronounce my name correctly, how can it manage my money properly?). In the worst case, the user could not recognise his/her name and he/she will not confirm it to the system producing a failed interaction. For these databases, the estimated stress assignment error rate is around 13.8%. It is not possible go through all the registers to correct these errors by hand. Because of its dynamic nature, this problem needs an automatic solution.

The paper is organised as follows. Section 2 describes the features used for calculating the Stress-Confidence Measure. In Section 3, we describe the method used for stress assignment and the proposed Discrimination Measure. Section 4 presents the experiments for the proposed method and a comparison with a 'k-Nearest Neighbours' approach. Finally, in Section 5 we review the conclusions of this work.

2. FEATURES FOR STRESS ASSIGNMENT

In Spanish, there is a direct relationship between word phonemes and word graphemes. Because of this, spelling and pronunciation features provide the same information for stress assignment. In this work (as proposed in [1][2]), the features considered for stress assignment were:

- *Number of vowels:* number of vowels in the name considered.
- *Vowel Position:* vowel position in the name considering only the vowels in the word.
- *Vowel Context:* set of letters from the vowel context. We will consider different context sizes.
- *Name End:* set of letters from the name ending. In this case, we will evaluate several sizes for the word ending.
- *Vowel:* the vowel under analysis.

3. STRESS ASSIGNMENT

The proper names are the most difficult words for stress assignment because of its great variability. In this situation, the patterns recognition methods have to deal with more difficulties. In our case, for assigning stress to a name, we first calculate a Stress-Confidence Measure for each vowel in the word, and then we assign the stress to the vowel (syllable) with the best score.

3.1 Stress Confidence Measure.

To calculate the Stress Confidence Measure for each vowel we combine the features described in section 2 using a Multi-Layer Perceptron (MLP). In the experiments we tested different alternatives for the inputs to the Multi-Layer Perceptron. In all cases we consider the number of vowels in the name, coded with 5 bits. We differentiate between names with 1, 2, 3, 4, or more than 4 vowels. The vowel position is coded with another 5 bits, 1st position, 2nd position,...,and 5th or higher position. For the rest of inputs we consider a window of letters, 5 or 7 depending on the experiments, coded with 30 bits per letter (in Spanish we use 30 different graphemes). These letters are obtained from the context (CXT) of the vowel considered and the name ending (NE). In Table 1 we present the different configurations considered. The hidden layer consisted of 50 neurons. This value permits a compromise between the MLP flexibility and the amount of data needed for the training set, even in the experiment with more inputs (when we considered 7 letters). Only one output node was used to model the stress-confidence. During weight estimation, a target value of 1 is assigned when the vowel is stressed and a value of 0 when it is not stressed. The weights calculation is performed with the Backpropagation algorithm [3][4]. The non-linear function considered in each perceptron is the hyperbolic tangent [5] and the learning coefficient is fixed to 0.5.

Different Input Configurations	
5 letters	2 CXT + VOWEL + 2 CXT + 0 NE
	1 CXT + VOWEL + 1 CXT + 2 NE
7 letters	3 CXT + VOWEL + 3 CXT + 0 NE
	2 CXT + VOWEL + 2 CXT + 2 NE
	1 CXT + VOWEL + 1 CXT + 4 NE

Table 1: Different configurations for the letter input, considering a window of 5 and 7 letters.

In Figure 1, we can see an example of input configuration. In this case we consider the number of vowels, vowel position and 5 letters, 3 from the vowel context (including the vowel to analyse) and 2 from the name ending. The number of binary inputs is 5+5+5x30= 160.

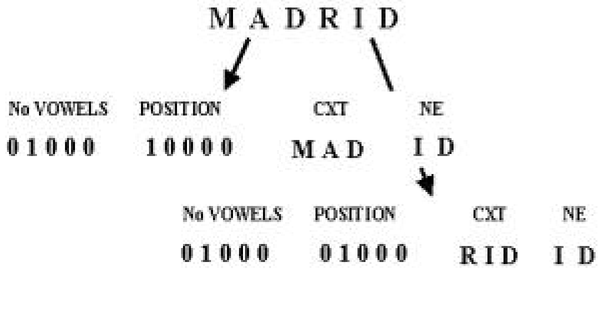

Figure 1: Example of inputs configuration for the two vowels of the name Madrid.

3.2 Discrimination Measure.

When we want to study different input configurations, it is necessary to analyse the power discrimination for each of them in order to foresee their behaviours in the classification task. Before describing the measure proposed, we will comment the two discrimination parameters considered:

- *Percentage of Ambiguous Input Patterns (PAIP)*: this is the percentage of cases, along the training set, where a same input pattern can produce different outputs (ambiguous pattern), i.e. we have vowels stressed and not stressed with the same input pattern.
- *Global Confusion Measure (GCM):* for each ambiguous pattern, we define a confusion measure (CM) given by the equation 1.

$$CM = \frac{100 - average(abs(D_{ij}))}{100} \qquad (1)$$

Where D_{ij} is the difference between the occurrence percentage (in the training set) of the classes i and j, for the ambiguous input considered. For example, if we have an ambiguous input pattern that produces 63.8% times a stressed vowel and the 36.2% a non-stressed vowel, then CM = 0.72. When we have similar percentages, $D_{ij} \rightarrow 0$, and we have the maximum confusion CM = 1, but when we have different percentages, $D_{ij} \rightarrow 100$, and the confusion tends to be zero (CM=0). A Global Confusion Measure can be obtained combining linearly the CMs from all the ambiguous patterns with the equation 2.

$$GCM = \sum_i W_i \times CM_i \qquad (2)$$

Where W_i is the percentage of ambiguous instances (examples belonging to an ambiguous pattern) that correspond to the ambiguous pattern i.

To calculate the Discrimination Measure (DM) for each input configuration, we combine these parameters following equation 3:

$$DM = 100 - (PAIP \times GCM) \qquad (3)$$

In Table 2, we can see these values for the different configurations that we have considered. In this case, we have not included the information from the number of vowels and the vowel position.

	Configurations	PAIP	GCM	DM
5 letters	2 cxt +V+2 cxt + 0 ne	17.1	0.35	94.0
	1 cxt +V+1 cxt + 2 ne	9.9	0.29	97.1
7 letters	3 cxt +V+3 cxt + 0 ne	3.6	0.40	98.6
	2 cxt +V+2 cxt + 2 ne	2.4	0.33	99.2
	1 cxt +V+1 cxt + 4 ne	2.9	0.25	99.2

Table 2: Discriminations measures for the different input configurations, considering 5 or 7 letters.

We can see how the letters from name ending are important for stress assignment. In both cases, with 5 or 7 letters, when we include name ending information the discrimination power of these configurations increases. It was a foreseeable result, as in Spanish there are some endings that are always stressed (e.g., "-ción" endings).

Using more letters (7 instead of 5) the discrimination is bigger. If we consider the information from the other two features: number of vowels and vowel position, the values for the Discrimination Measures are the following:

	Configurations	PAIP	GCM	DM
5 letters	2 cxt +V+2 cxt + 0 ne	7.6	0.47	96.4
	1 cxt +V+1 cxt + 2 ne	2.6	0.29	99.2
7 letters	3 cxt +V+3 cxt + 0 ne	1.7	0.59	99.0
	2 cxt +V+2 cxt + 2 ne	1.5	0.46	99.3
	1 cxt +V+1 cxt + 4 ne	1.7	0.39	99.3

Table 3: Discriminations measures for the different input configurations, considering 5 or 7 letters, the number of vowels and the vowel position.

In this case, the introduction of the features number of vowels and vowel position has caused that the 5 letters configuration "1 ctx + V + 1 ctx + 2 ne" has more discrimination power that the 7 letters configuration "3 ctx + V + 3 ctx + 0 ne". This fact shows that the new features considered and the long context information are redundant, i.e. they have similar information.

4. EXPERIMENTS

For the experiments we have considered a list of 48,396 different proper names taken from the most common Spanish proper names directory obtained in [6] (first-names and surnames). From this set of names we have taken at random 4,838 of them for testing, leaving 38,719 for training and 4,839 for evaluating. In this situation, we have 118,792 patterns to train the MLP (vowels in the name training list), 14,949 for evaluation and 14,876 for testing.

4.1 Results.

Considering the different input configurations the results obtained are shown in Table 4:

	Configurations	Correctly stressed rate (%)
5 letters	2 cxt +V+2 cxt + 0 ne	93.5
	1 cxt +V+1 cxt + 2 ne	94.5
7 letters	3 cxt +V+3 cxt + 0 ne	94.1
	2 cxt +V+2 cxt + 2 ne	94.9
	1 cxt +V+1 cxt + 4 ne	94.9

Table 4: Correctly stressed name rates for the different input configurations, considering a window of 5 and 7 letters.

As we can see, the results confirm the main conclusions mentioned in the previous section. We can see how a bigger number of input letters produces better results except for the case "3 ctx+V+3 ctx + 0 ne" where the performance is worse as we can see in Table 4. The information from the name ending is also very important for stress assignment. It is necessary to get a compromise between the context and name-ending information. We can conclude that there is an important relationship between these results and the values calculated for the Discrimination Measure that we propose.

4.2 Comparison with the 'k-Nearest Neighbours' approach

Now, we are going to compare the results obtained with the MLP with the results obtained using a Memory-Based Learning approach (k-Nearest Neighbours). In this technique, the training lies in storing in memory all feature patterns from the training set, with its solution (stressed or non-stressed). In the testing stage, for each pattern analysed, we calculate a distance D(X,Y) between it and all the training patterns kept in memory. The solution will be the most frequent solution from the set of the K closest training patterns [7]. In our case K=1, so the solution for the new pattern will be the solution stored for the closest training pattern. If a pattern is associated with more than one category in the training set (i.e. the pattern is ambiguous), the distribution of patterns over the different categories is kept, and the most frequently occurring category is selected when the ambiguous pattern is used to classify. To calculate the distance between patterns we considered the *Similarity Metric:*

$$D(X,Y) = \sum_{i=1}^{N} d(x_i, y_i) \qquad (4)$$

where N is the number of features (dimension of the patterns X and Y) and x_i and y_i are the values of feature i in the patterns X and Y. The distance $d(x_i, y_i)$ is calculated using the equation 5:

$$d(x_i, y_i) = \begin{cases} 0 & if \;\; x_i = y_i \\ 1 & if \;\; x_i \neq y_i \end{cases} \qquad (5)$$

With this metric (equation 4) all the features have the same importance. We can add more knowledge analysing the behaviour of each feature in the training set. The features with higher discrimination power should contribute with a larger weight to the distance. So, we considered a weighted distance, as in [8]:

$$D(X,Y)=\sum_{i=1}^{n} W_i\, d(x_i,y_i) \qquad (6)$$

To calculate the weights W_i, we use the *Information Gain (IG)* [9][10]. The IG for a specific feature is the difference between the Information Entropy (IE) of the training set, and the average Information Entropy along the different training subsets obtained dividing the training set according to the different values of the feature considered. If we analyse a continuous feature, we have to define a finite number of ranges.

The Information Entropy for a set is calculated using the equation 7.

$$IE=-\sum_{p_i} p_i \log p_i \qquad (7)$$

where p_i is the probability for the class i. These probabilities are estimated considering the class frequency in the set. In our case we only have two classes: stressed and non-stressed vowels. For calculating the average Information Entropy for a feature, we weighed the contribution from each subset with its size in relation to the whole training set. In our case, the weights W_i are equal to the Information Gain for feature i.

Considering the configuration "2 ctx+V+2 ctx + 2 ne", we have evaluated the Nearest Neighbour approach described above over the same training and testing sets. The correctly stressed name rate obtained is 93.1%. The MLP produces better results (94.9%) and the processing time is lower.

5. CONCLUSIONS

We have presented an approach based on a Multi-Layer Perceptron (MLP) for Stress Assignment in Spanish Proper Names. In the paper we present and analyse different alternatives for the inputs to the Multi-Layer Perceptron. We propose a *Discrimination Measure* to analyse the discrimination power for the different input configurations and we validate this measure presenting the results obtained in each case. For the best combination we obtained a 94.9% proper names correctly stressed (5.1% error rate) with a processing time of 8.4 ms for stress assignment. These results are compared favourably to similar experiments using a *Memory Based Learning* approach (k-Nearest Neighbours). So, we can conclude that the Multi-Layer Perceptron is a good tool to combine different sources of word stress information allowing us to reduce 63.0% the stress assignment error rate (from 13.8% to 5.1%)

In relation to the features considered, we can point out that the vowel context and the word-ending information are very useful for stress assignment and it is important to get a compromise between these sources of information. On the other hand, the number of vowels and the vowel position are the features with the lowest discrimination power.

6. REFERENCES

1. G. Durieux “Analogical Modelling Of Main Stress Assignment in Dutch Simplex Words”. Daelemans and Powers (Eds.) Background and Experiments in Machine Learning of Natural Language. Proceedings First SHOE Workshop. Tilburg: ITK, pp. 197-204, 1992.
2. S. Gillis, W. Daelemans, G. Durieux, A. Van den Bosch. "Learnability and Markedness: Dutch Stress Assignment". Proceeding of the Fifteenth Annual Conference of the Cognitive Science Society; pp.452-457, 1993.
3. D. Rumelhart and J. L. McClelland. "Parallel Distributed Processing: Vol. 1: Foundations". MIT Press, 1986.
4. B. Widrow. R. G. Winter, and R. Baxter. "Layered neural nets for pattern recognition". IEEE Trans ASSP, Vol. 36. No. 7, pp. 1109-1118.
5. R.M. Golden. "Mathematical Methods for Neural Network Analysis and Design". MIT Press, 1996.
6. The ONOMASTICA Consortium, “The ONOMASTICA interlanguage pronunciation lexicon”. Proceedings of Eurospeech'95, Madrid, Vol. I, pp. 829-832, 1995.
7. J. Zavrel and W. Daelemans, "Memory-based learning: Using similarity for smoothing". Computing and the Humanities. Special Issued on Senseval, 1997.
8. S. Cost, and S. Salzberg. "A weighted nearest neighbour algorithm for learning with symbolic features". Machine Learning 10: pp. 57-58, 1993.
9. W. Daelemans and A. van de Bosch "Generalization Performance of Backpropagation Learning on a Sylabification Task". Connectionism and Natural Language Processing Third Twente Workshop on Language Technology: pp. 27-38, 1992
10. W. Daelemans, S. Gillis, G. Durieux, A. Van den Bosch "Learnability and Markedness in Data-Driven Acquisition of Stress". Computational Phonology. Edinburg Working Papers in Cognitive Science 8: pp.157-178, 1993.