Evaluation of Machine Learning Approaches to Portuguese Part-of-Speech Prediction

D. C. Cavalieri[1], S. E. Palazuelos-Cagigas[2], T. F. Bastos-Filho[1] and M. Sarcinelli-Filho[1]

[1] Departamento de Engenharia Elétrica
Universidade Federal do Espírito Santo - UFES
Vitória, Espírito Santo, Brazil
[2] Departamento de Electrónica, Escuela Politécnica Superior
Universidad de Alcalá - UAH
Alcalá de Henares, Madrid, España
[1]{daniel, tfbastos, mario.sarcinelli}@ele.ufes.br
[2]sira@depeca.uah.es

Abstract. The prediction of the part-of-speech of the next word is a technique commonly used to improve prediction systems and, therefore, to reduce the amount of keystrokes that a user needs to enter text in a computer. In this paper, we have adapted and implemented a number of existing part-of-speech prediction algorithms, such as the Statistical Part-of-Speech Model, Artificial Neural Networks, Naive Bayes predictors, Support Vector Machines and Linear Regression predictors, and introduced new ones, such as meta-learning ranking to select a candidate algorithm, and a fusion algorithm to combine their results to perform relevance assessments on the Portuguese part-of-speech prediction and, therefore, to the Portuguese word prediction. Their effectiveness is illustrated by comparing the results obtained by each prediction method and the ones obtained without grammatical information.

1 Introduction

The neuromotor disability is one of the main causes of problems associated to the speech and the communication of people with disabilities and the environment around them. Since many of these people use computers for personal communication, they experiment frustration because their word/phrase composition rate is too low to keep a normal conversation. Thus, word prediction methods are commonly used to assist them in the task of communication or text writing.

In this context, evaluation of word prediction methods plays a crucial role in the development of this kind of systems, since it gives rise to objective performance measurements, allowing comparisons between different techniques and accurate measurements of the improvements (if any) due to new theoretical or algorithmic approaches.

In this work we evaluated techniques for Portuguese *part-of-speech* (*pos*) prediction, to improve the word prediction system called PREDWIN, already developed in [9] for Spanish language.

2 Related Work

During the last few years, we have witnessed considerable activity in the field of Natural Language Processing (NLP) system assessment. Most of the existing word-prediction systems exploit statistical prediction algorithms with word frequency information [9]. Some authors [7,4] have developed communication models to reduce the number of keys representing the English alphabet that attempt to help people with disabilities. These systems use an n-gram model for characters, created from text samples and from the user's input. The models are stored in a special tree structure that allows partial matches between context and model to find the letters rapidly.

In order to improve the accuracy of predictors, some of them consider a larger context and use word sequence statistics, such as word bigram or trigram models. Other authors [8,9] use word n-gram statistics along with optimal customization to each user's vocabulary in order to suggest appropriate words. These predictors achieved between 30% and 53.1% keystroke savings, depending on the type of text, adaptation to the user, and the type of frequency information used. In all the experiments, the size of the prediction list was between 5 and 10 words.

Other way to improve the accuracy of prediction systems is to include syntactic information, by adding word category statistics or grammar rules. The goal of syntactic prediction is to ensure that the system does not suggest to the user grammatically incorrect words. Some of these systems consider the frequency of different length sequences of word *pos* as syntactic information [9,5,2,12], while other ones use a parser to build the syntactic structure of the whole sentence [14,6]. These systems have provided an improvement between 0.5% and 2% on the keystroke saved, depending on the text and the number of predictions offered.

3 Portuguese *POS* Prediction

In this paper, the results of the following seven *pos* prediction methods are presented: Statistical *pos* models (N-POS); Artificial Neural Networks (ANN) trained with sequences of 2, 3 and 4 Portuguese *pos*; Support Vector Machines (SVM), Logistic Regression Models (LR) and a Naive Bayes Classifier (NB) trained with sequences of 2, 3, 4 and 5 Portuguese *pos*, all described in [3]. In addition to that, we also briefly describe and provide results about a meta-learning strategy used for algorithm selection and an algorithm developed to make the fusion of all models.

The sequences of *pos* were generated from a set of newspaper articles contained in the Portuguese corpus called CHAVE [10], whose syntactic information was obtained using an automatic categorizer named PALAVRAS [1].

3.1 Meta-Learning for Algorithm Selection

In general, there are three options to generate the output of the meta-learner. The first one is to select a single learning algorithm, i.e. to select the algorithm

that is expected to produce the best model for the dataset. The second is to select a sub-group of learning algorithms, including not only the best algorithm but also the algorithms that are not significantly worse that the best one. The third possibility is to rank the learning algorithms according to their performance. This performance can be measured through the properties of the class probability distribution predicted by the algorithms for a given example: number of classes, number of features, ratio of examples to features, average class entropy, signal to noise ratio, etc [13].

In this work we have used two properties of the class distribution to select an algorithm: first, the highest probability of the predicted class, given by

$$maxprob = \max_{k=1}^{K} p(C_k), \tag{1}$$

Next, the entropy of the class probability distribution, given by

$$entropy = -\beta \cdot \sum_{k=1}^{K} p(C_k) \log_2(p(C_k)), \tag{2}$$

where K is the number of Portuguese *pos* and $p(C_k)$ is the probability value of each class, i. e., the k-th element denotes the probability that the example $\mathbf{x}$ belongs to a class C_k as estimated by the algorithms used. β is a dimensional adjustment constant.

Both, the entropy and the maximum probability of a probability distribution, can be interpreted as estimations of the confidence of the model in its prediction. If the probability distribution returned is highly spread, the maximum probability will be low and the entropy will be high, indicating the model is not very confident in its prediction. On the other hand, a high probability and a low entropy indicate that the model is confident in its prediction [11]. Considering that, we can measure the distance between those two properties using the Euclidean distance, given by

$$dist = \sqrt{maxprob^2 + entropy^2}. \tag{3}$$

Thus, the higher value of *dist* selects the best candidate algorithm to be used.

3.2 Algorithm Fusion

A key idea in this work is that if one algorithm is sometimes superior to and sometimes inferior to the others, then it may be possible to combine them to achieve better performance than using each one in isolation.

The algorithm fusion works as follows. After acquiring the class probability distribution of each algorithm applied to an example feature vector $\mathbf{x}$, a vector $\mathbf{P'} = \left[p'(C_1)\, p'(C_2) \cdots p'(C_K)\right]$ with the new values of class probability distribution is generated according to the following equation:

$$p'(C_k) = \sum_{i=1}^{n} p_i(C_k), \ \forall k, \tag{4}$$

where n is the number of algorithms used.

Thus, a normalization factor C is needed and it is calculated by:

$$C = \sum_{k=1}^{K} p'(C_k), \tag{5}$$

Finally, a new class probability distribution is obtained dividing 4 by 5:

$$\mathbf{P}_{(fusion)} = \frac{p'(C_k)}{C}, \ \forall k. \tag{6}$$

4 Results

In this work, 76 Portuguese *pos* were chosen, being formed by 10 basic word categories (noun, verb, adjective, etc.), adding some possible inflections features (gender, number, grade), some punctuation marks (period, comma, dash, etc.), and some verb conjugation information in each *pos* that could accept it.

Two main experiments have been carried out in order to evaluate the algorithms used here. In the first one, a list of 135,482 Portuguese word *pos*, extracted from the corpus, was used in the training and validation stage of each method. In the evaluation stage, a list of 73,465 Portuguese word *pos* and punctuation, extracted from the same corpus was used. The texts used to train and test do not overlap. In these tests we evaluate whether the word *pos* obtained from the class probability distribution of each algorithm applied is the right one or not. The evaluation results are listed in Table 1.

Next, the methods presented in Table 1 were incorporated on the general word prediction system PREDWIN and a final test was performed. The test text was found on the Internet, contained 84,174 words and punctuation marks, and a total of 449,194 keystrokes are needed to write it without any prediction method. As baseline results for the performance comparison, an *unigram* word model without any *pos* prediction method that only requires the written part (letters) of the current word was used. Table 2 contains the final results for the word prediction evaluation.

The dashes in Tables 1 and 2 correspond with algorithms that are still under development (training and experimentation stage) due to the computational resources they require (memory and processing time).

Table 1. Results obtained in the Portuguese *pos* prediction evaluation using 2, 3, 4 and 5 *pos* sequence length.

	Accuracy(%)			
Method	2	3	4	5
RNA	24.50	28.55	29.15	-
SVM	23.73	25.07	25.80	25.84
LR	24.23	25.39	26.38	26.69
Bayes	24.47	24.04	23.68	23.67
Meta-Learning Ranking	25.03	28.75	30.45	31.12
Algorithm Fusion	24.43	25.02	25.75	26.01

Table 2. Results obtained in the Portuguese word prediction evaluation using 2, 3, 4 and 5 *pos* sequence length.

	Relative Improvement (%)				Processing Time (ms/word)			
Method	2	3	4	5	2	3	4	5
n-gram	Baseline (32.31%)				3.40			
N-POS	0.92	1.18	-	-	20.73	20.35	-	-
RNA	2.47	2.71	2.57	-	37.32	53.11	76.18	-
SVM	2.69	2.87	2.90	3.15	232.64	241.99	230.93	180.17
LR	2.81	3.03	3.08	3.10	8.49	8.61	8.62	8.52
Bayes	2.92	2.82	2.81	2.80	9.14	9.26	9.24	8.93
Meta-Learning Ranking	-0.46	-0.25	-0.16	-0.02	320.55	362.03	388.32	355.48
Algorithm Fusion	3.27	3.42	3.38	3.40	189.57	213.23	184.52	375.21

5 Conclusions

Our goal in this work was to investigate how we can improve the performance of Portuguese word prediction by including *pos* information on existing prediction systems, such as PREDWIN. We have adapted and implemented a number of existing *pos* prediction algorithms, such as N-POS, RNA, Bayes, SVM and LR predictors. We have also proposed a meta-learning ranking to select a candidate algorithm, and a fusion algorithm trying to combine their results to perform relevance assessments on the Portuguese *pos* prediction and, therefore, to the Portuguese word prediction.

In other experiment, we observed that all the methods evaluated here showed a higher performance than methods based only on the frequency of the categories (bipos and tripos). When compared with the prediction method without grammatical information (n-gram), the proposed approach based on algorithm fusion shows an increase of up to 3.42% in the number of keystrokes saved by the user. Also, this approach can be used to include other prediction algorithms providing complementary information.

The Portuguese *pos* prediction algorithms gave an improvement between 0.92% and 3.42% in the keystrokes saved, which is similar or even better than the improvement values found by the best English prediction systems [9,2,12,14,6], that also use syntactic information. Thus, this confirms the validity of the *pos* prediction algorithms evaluated.

The processing time needed for each method has also been studied, and is showed in Table 2. It is low when compared with the time needed for the user with disabilities to compose a word/phrase using a communication interface.

Acknowledgments

We would like to thank CAPES/Brazil and MECD-AGU/Spain (Project 150/07) and FACITEC/Brazil for the financial support given to this work.

References

1. Bick, E.: The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Ph.D. thesis, Aarhus University, Aarhus, Denmark (November 2000)

2. Booth L., Morris C., R.I.W., F., N.A.: Using a syntactic word predictor with language impaired young people. In: Murphy, H. (ed.) In Proceedings of the California State University, Northridge (CSUN), 7th Annual Conference on Technology and Persons with Disabilities. pp. 57–61. California State University, Northridge, Los Angeles, California, USA (March 1992)
3. Cavalieri, D., Palazuelos, S., Bastos, T., Filho, M.S.: Evaluation of machine learning approaches to portuguese part-of-speech prediction (written in portuguese). In: SBAI 2009 (April 2009)
4. Darragh, J.J., Witten, I.H., James, M.L.: The reactive keyboard: A predictive typing aid. Computer 23(11), 41–49 (1990)
5. Fazly, A., Hirst, G.: Testing the efficacy of part-of-speech information in word completion. In: Proceedings of the 10 th Conference of the European Chapter of the Association for Computational Linguistics. pp. 9–16 (2003)
6. Garay-Vitoria, N., Gonzales-Abascal, J.: Intelligent word prediction to enhance text input rate (a syntactic analysis based word prediction aid for people with severe motor speech disability). In: Annual International Conference on Intelligent User Interfaces. pp. 241-247 (1997)
7. Mackenzie, I.S., Kober, H., Smith, D., Jones, T., Skepner, E.: Letterwise: Prefix-based disambiguation for mobile text input. In: In Proceedings of the 14th annual ACM Symposium on User Interface software and technology. pp. 111–120. ACM (2001)
8. Nantais T., Shein F., J.M.: Efficacy of the word prediction algorithm in wordq(tm). In: Simpson, R. (ed.) Proceedings of the RESNA 2001 Annual Conference: The AT Odyssey Continues. pp. 77–79. RESNA Press Rehabilitation Engineering and Assistive Technology Society of North America, Arlington, VA (September 2001)
9. Palazuelos-Cagigas, S.E.: Contribution to word prediction in Spanish and its integration in technical aids for people with physical disabilities. Ph.D. thesis, Universidad de Alcalá de Henares, Alcalá de Henares, Madrid, Spain (2001)
10. Santos, D., Rocha, P.: The key to the first clef in portuguese: Topics, questions and answers in chave. In: 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004. pp. 821-832. Bath, UK (September 15-17 2004)
11. Todorovski, L., Dzeroski, S.: Combining multiple models with meta decision trees. In: PKDD '00: Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery. pp. 54–64. Springer-Verlag, London, UK (2000)
12. VanDyke, J.A.: A syntactic predictor to enhance communication for disable users. Technical Report 92-03, Department of Computer and Information Sciences, University of Delaware (1991)
13. Vilalta, R., Drissi, Y.: A perspective view and survey of meta-learning. Artificial Intelligence Review 18, 77–95 (2002)
14. Wood, M.E.J.: Syntactic PreProcessing in Single-Word Prediction for Disabled People. Ph.D. thesis, Department of Computer Science, University of Bristol (June 1996)