Combination of Language Models for Word Prediction: An Exponential Approach

Daniel C. Cavalieri, Sira E. Palazuelos-Cagigas, Teodiano F. Bastos-Filho, and Mário Sarcinelli-Filho

***Abstract*—This paper proposes an exponential interpolation to merge a part-of-speech-based language model and a word-based n-gram language model to accomplish word prediction tasks. In order to find a set of mathematical equations to properly describe the language modeling, a model based on partial differential equations is proposed. With the appropriate initial conditions, it was found an interpolation model similar to the traditional maximum entropy language model. Improvements in keystroke saved and perplexity over the word-based n-gram language model and two other traditional interpolation models is obtained, considering three different languages. The proposed interpolation model also provides additional improvement in hit rate parameter.**

***Index Terms*—Natural language processing, word prediction, combination of language models.**

I. Introduction

WORD prediction systems (WP) were developed as a communication aid method, in order to increase message composition rate for people with severe motor and speech disabilities [1]. Nowadays, text prediction methods suitably integrated into user interfaces can benefit anyone trying to produce text messages or commands [2]. Generally, prediction refers to those systems that guess which letters, words, or phrases are likely to follow a given segment of a text [3]. In [4] and [5] the word prediction system is considered as an important task within the context of Natural Language Processing (NLP), in which the goal is to predict the correct word given a particular context. In all the cases, the main goal of these systems is to increase the keystroke saving (KSS), which is the percentage of keystrokes that the user saves by using word prediction systems, besides ensuring a good quality of the produced text.

There are several WP systems that have been and are being developed using distinct methods for different languages [3], [6]. Traditionally, these systems have been based on statistical n-gram language modeling. Recently, more sophisticated language models have been developed in order to improve the performance of these traditional language models [7]. In many cases, each one of these language models explores and captures separately specific phenomena of natural language. Here, a question that naturally arises is: how to build more powerful and complex language models, capable of integrating all language components such as syntactic, semantic and morphological structures?

To answer this question, the more efficient method is to combine them in some optimal sense [8]. A simple and widely combining method is the linear interpolation, which takes into account a weighted sum of the probabilities given by the component language models. Normally, it is used to add a Part-of-Speech (POS) cache-component to a word n-gram model [9], taking into account the semantic structure of the language [10], or both POS and semantic structures [11]. Nonetheless, according to [12] even if the perplexity of the linear combined model is minimized, this type of methodology does not guarantee optimal use of the different information sources. This way, [13] proposes a method based on the Latent Maximum Entropy Principle, which extends the basic principle of maximum entropy proposed in [14], incorporating a hidden dependence structure. Distinct from linear interpolation, this approach generates probabilistic models capable of capturing all information from different sources, but with computational limitations in the estimation of the model parameters. In order to improve this, [13] also proposed a methodology based on Directed Markov Random Fields, first developed in [15]. With this model, the authors were able to combine a word trigram model, a probabilistic model based on context-free grammar and a probabilistic model based on latent semantic analysis. In the same way, [16] proposes the Sentic Computing, which involves the use of Artificial Intelligence and Semantic Web techniques to represent knowledge, mathematics to perform tasks such as graphical representations and dimensional reduction, linguistics to discourse analysis, psychology to cognitive and affective modeling, sociology to better understand the dynamics and influence of social networks and, finally, ethics to understand the relationships between nature of mind and creating emotional machines. In this sense, Figure 1 shows an envisioned evolution proposed by [16] of Natural Language Processing (NLP) research through the main elements of natural language: *syntactic*, *semantic* and *pragmatic*. However, it is not clear in their work how to combine all this information to be used in pratical systems. Then, the idea of using partial differential equations, proposed in this work, seems interesting as we can mathematically model the *best path* to combine all the main elements in the natural language processing and, then, improve the Sentic Computing methodology.

Manuscript received February, 2015. This work was supported by the Spanish Ministry of Science and Innovation under projects VISNU (Ref. TIN2009-08984) and SD-TEAM (Ref. TIN2008-06856-C05-05), by CAM-UAH under FUVA project (CCG10-UAH/TIC-5988) and by CAPES/Brazil under grant 150/07.

T. F. Bastos-Filho and M. Sarcinelli-Filho are with the Department of Electrical Engineering, Universidade Federal do Espírito Santo, Vitória, 29075-910 Brazil (e-mail: tfbastos@ele.ufes.br; mario.sarcinelli@ufes.br).

D. C. Cavalieri is with the Department of Control and Automation Engineering,Instituto Federal do Espírito Santo, Serra, 29173-087 Brazil (e-mail: daniel.cavalieri@ifes.edu.br.

S. E. Palazuelos-Cagigas is with the Department of Electrical Engineering, Universidad de Alcalá, Alcalá de Henares, Madrid 28801, Spain (e-mail: sira@depeca.uah.es).

Fig. 1. Envisioned evolution of NLP research (extracted from [16]).

As it can be seen, the above-mentioned language models reached a high level of mathematical (and computational) complexity, despite the efforts dedicated to minimize it. Such models are widely used in applications such as automatic speech recognition (ASR) and machine translations (MT). However, when dealing with word prediction, it is recognized that the syntactic structure of the language plays a key role, many times a primary one, in the composition of the language model. Thus, this work is motivated by the assumption that n-gram models could be more effective in WP, as the main model, performing its combination with POS-based language models, which provides additional information, and proposes a novel exponential approach, theoretically based on partial differential equations to combine them. As in many natural processes, once determined such differential equations, that characterize a particular system, it is possible to extract relevant information about them. Figure 2 gives a general overview of the proposed methodology.

Fig. 2. General overview of the proposed methodology.

Our hypothesis was tested using English, Portuguese and Spanish. For each language, a word n-gram model (with $n = 4$) and a m-POS language model (with $m = 3$) formed by the linear interpolation of three different POS-based language models was built, with each linear weight coefficient (β) based on the Area Under a Receiver Operating Characteristic Curve (AUC). To corroborate the methodology, results obtained using the proposed interpolation model were compared with the linear and geometric interpolations, respectively described as

$$\begin{aligned} P^{(linear)}_{interpolation}(w_i|w_{i-(n-1)}, t_{i-(m-1)}) = \\ \alpha\cdot[P_{n-gram}(w_i|w_{i-(n-1)})]+ \\ +(1-\alpha)\cdot[P_{m-POS}(w_i|w_{i-(m-1)})] \end{aligned} \quad (1)$$

and

$$\begin{aligned} P^{(geometric)}_{interpolation}(w_i|w_{i-(n-1)}, t_{i-(m-1)}) = \\ [P_{n-gram}(w_i|w_{i-(n-1)})]^{\alpha}\cdot \\ \cdot[P_{m-POS}(w_i|w_{i-(m-1)})]^{(1-\alpha)}. \end{aligned} \quad (2)$$

Some studies are presented here as frameworks to compare our proposal, in terms of performance. Among them, it is worth mentioning the word prediction system based on word classes and topics presented in [17], the language model proposed in [18], defined as a linear combination of a word n-gram model and a POS-based language model, both for English, and the statistical POS-based language model proposed in [19] for Spanish. It is also important to mention the hybrid model initially proposed in [20] and used in [21] for French, constituted by a geometric interpolation of a Topic-based model and a Word-based model. Improvements for each language model in terms of *keystroke saved* (KSS) and *perplexity* (PP) are shown here, in addition to a demonstration of an improvement in the number of words predicted before the user typed any letter (known as *hit rate* - HR).

The main contributions of this work can be described as a proposal of:

1) a general mathematical model, based on partial differential equations, that represents a WP language model;
2) a novel interpolation method based on an *natural* exponential interpolation of a word-based n-gram model and a POS-based language model.

The rest of the paper is organized as follows: section II gives a general overview of the Word-based model and the POS-based model. In Section III our proposed interpolation method to combine these language models is addressed. Section IV reports the outcome of experimental evaluations conducted using a WP system in English, Portuguese and Spanish. Finally, Section V shows conclusions and outlines for future works.

II. Language Models

A. Word-based Language Model

In word prediction, a statistical language model tries to predict the next word based on the history of previous words. This idea of word prediction is formalized by probabilistic models called n-gram models, which in turn predict the next word from the $n-1$ previous words. In its simplest version,

the unigram model only considers the absolute frequency of the word. When using this model, at each moment the most frequent words that begin with written letters of the word in progress are predicted. When considering the sequence of words, or the probability that each word follows the previous words, there exist bigram, trigram, ..., n-gram. Otherwise, suppose a sentence in which the sequence of words given so far is

$$w_{i-(n-1)} \cdots \ w_{i-2} \ w_{i-1} \ cw_i,$$

where w_{i-n-1} are the $n-1$ previous words and cw_i is the current word prefix typed by the user. Thus, the bigram model ($n=2$) is given by

$$P_{bigram}(w_i|w_{i-1}) = \frac{F(w_{i-1}w_i)}{\sum_{w_i} F(w_{i-1}w_i)}. \quad (3)$$

Such equation can be simplified, since the sum of all the bigrams beginning with the w_{i-1} must be equal to the count of the word unigram. Then,

$$P_{bigram}(w_i|w_{i-1}) = \frac{F(w_{i-1}w_i)}{F(w_{i-1})}, \quad (4)$$

which can be easily extended to the n-gram model, or

$$P_{n-gram}(w_i|w_{i-(n-1)}) = \frac{F(w_{i-(n-1)} \dots w_i)}{F(w_{i-(n-1)} \dots w_{i-1})}, \quad (5)$$

where $F(w_{i-(n-1)} \dots w_i)$ and $F(w_{i-(n-1)} \cdots w_{i-1})$ are the frequencies of the n-th e $(n-1)$-th previous word sequences, respectively.

One disadvantage of the word n-gram language model is its large number of parameters. Another disadvantage of the word n-gram language model is its high dependence on the discourse domain when measuring perplexity on a set of different texts belonging to (or outside of) the linguistic domain of the training *corpus*.

B. POS-based Language Model

A solution to overcome the data sparseness problem and reduce the dependence on the discourse domain consists of grouping words together into equivalence word classes (or POS in our case), instead of those of individual words. In this context, before detailing the POS-based language model itself, it is interesting to (briefly) describe the technique used to determine the POS tagset used in this paper.

As showed in [22], the use of some major POS (noun, verb, adjective, etc) along with some inflections like gender (masculine, feminine, neuter), number (singular, plural, neuter) and person (1st, 2nd, 3rd, 1st/3rd) can generate accurately POS-based word predictors with a relatively low speed list of predicted words. Thus, an initial POS tagset was first derived by selecting the most functional POS tags correspondent to English, Spanish and Portuguese. Even showing a relatively low number of POS tags, this work also used the methodology developed in [23] to further reduce the number of POS tags in Spanish and Portuguese. Table I shows the POS tagset and the morphological analyzer used to each language.

By modeling the language with POS tags, the system predicts the next POS tag to be produced in the current sentence and narrows the amount of possible next words when each letter of the word is entered. In other words, a syntactic predictor has access to the following sequence of words and POS tags to predict the current word:

$$\cdots \ w_{i-2}/t_{i-2} \ w_{i-1}/t_{i-1} \ cw_i,$$

where t_{i-2} and t_{i-1} are the POS tags of the previous words w_{i-2} and w_{i-1}, respectively, and cw_i is the current word prefix typed by the user. The algorithm predicts words starting by cw_i.

There are different methods for incorporating the statistical POS tag information into the word predictor [18]. As in [19], the syntactic predictor (with $m=2$) was estimated by

$$P_{2-POS}(w_i|w_{i-1}) = \sum_{r=1}^{|T(w_{i-1})|} \sum_{s=1}^{|T(w_i)|} P(t_i^s|t_{i-1}^r) \cdot \cdot P(w_i|t_i^s) \cdot P(t_{i-1}^r|w_{i-1}), \quad (6)$$

where $|T(w_{i-1})|$ is the total number of POS tags that may be assigned to the previous word w_{i-1} and $|T(w_i)|$ is the total number of POS tags that may be assigned to the current word w_i. $P(t_i^s|t_{i-1}^r)$ is the bigram POS tag probability (the probability of t_i^s being t_{i-1}^r the tag of the previous word), and $P(w_i|t_i^s)$ is the conditional probability of the word w_i given t_i^s as its POS tag. $P(t_{i-1}^r|w_{i-1})$ is the conditional probability of the previous word w_{i-1} to be tagged with its rth tag t_{i-1}^r.

This method can be extended to include in the prediction as many previous words as desired. The current system considers a maximum of two previous words in the prediction ($3-POS$ model), or

$$P_{3-POS}(w_i|w_{i-1}, w_{i-2}) = \sum_{p=1}^{|T(w_{i-2})|} \sum_{r=1}^{|T(w_{i-1})|} \sum_{s=1}^{|T(w_i)|} P(t_i^s|t_{i-1}^r t_{i-2}^p) \cdot \cdot P(w_i|t_i^s) \cdot P(t_{i-1}^r|w_{i-1}) \cdot P(t_{i-2}^p|w_{i-2}), \quad (7)$$

where $|T(w_{i-2})|$ is the total number of POS tags that may be assigned to the previous word w_{i-2}.

There is an important difference concerning the POS-based language model used within this paper and the ones used in works like [18]. The difference lies in the calculation of the POS tag probabilities $P(t_i^s|t_{i-1}^r)$ and $P(t_i^s|t_{i-1}^r t_{i-2}^p)$. Traditionally, these probabilities are calculated using the frequencies

TABLE I
POS TAGSET TO PORTUGUESE, SPANISH AND ENGLISH.

	Portuguese	**Spanish**	**English**
Morphological Analyzer	PALAVRAS[24]	HISPAL[25]	ENGCG[26]
Initial Number of POS tags	259	282	129
Set of Functional POS tags	71	82	54
Reduced POS tagset using [23]	66	66	54

of the previous word classes, or

$$P(t_i|t_{i-(m-1)}) = \frac{F(t_{i-(m-1)} \cdots t_i)}{F(t_{i-(m-1)} \cdots t_{i-1})}. \quad (8)$$

In this work, a traditional statistical POS-based language model (equation 8), a Logistic Regression POS-based language model and a Naive Bayes POS-based language model are combined using the area under ROC curve (AUC). This methodology was proposed based on the work presented by [27], which propose different POS-based language models to Brazilian Portuguese.

ROC curve[1] is a technique for visualizing, organizing and selecting classifiers based on their performance. To compare classifiers it is possible to reduce the ROC performance to a single scalar value representing the expected performance. A common method is to calculate the AUC, which has an important statistical property: the AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance.

The AUC is usually estimated in the same way as the error rate, and to classify the accuracy of a classifier using this measure, the following equivalence is adopted:

- 90% 100%: excellent;
- 80% 90%: very good;
- 70% 80%: good;
- 60% 70%: fair;
- 50% 60%: poor;
- $< 50\%$: fail.

This way, a classifier that obtains an AUC of 86%, for example, will be considered a very good classifier, with a score of $\beta = 0.86$. In the same way, if a classifier obtains an AUC less than 50% it will fail and the score will be approximated to $\beta = 0$.

ROC analysis and AUC are commonly employed in two-class problems and with more than two classes, as in our case, the situation becomes much more complex. For handling this problem, different AUCs were calculated, one for each POS tag, using the *one-against-all* method. Thus, for a set of $|T(w_i)|$ POS tags, one can have $\boldsymbol{\beta}_{(.)} = [\beta_1, \beta_2, \ldots \beta_{|T(w_i)|}]$ and the combined m-POS language model given by

$$\begin{aligned} P(t_i|t_{i-(m-1)})^{(combined)} = \\ \boldsymbol{\beta}_{(1)} \cdot P(t_i|t_{i-(m-1)})^{(1)} + \boldsymbol{\beta}_{(2)} \cdot P(t_i|t_{i-(m-1)})^{(2)} + \\ + \boldsymbol{\beta}_{(3)} \cdot P(t_i|t_{i-(m-1)})^{(3)}, \end{aligned} \quad (9)$$

where (1) represents the traditional statistical POS-based language model, (2) represents the LR POS-based language model, and (3) represents the Naive Bayes POS-based language model.

III. Proposed Method

A. Mathematical Formulation

Before presenting the interpolation model proposed in this paper, the work presented in [21], that uses a model based on geometric interpolation, is discussed in detail, to better understand the relationship between the independent models and the mathematical model merging them.

Mathematical models seek to explain quantitatively and qualitatively natural phenomena, and are usually described through differential equations to describe the dynamic evolution of systems [29]. By solving such equations it is possible to extract relevant information about such systems and possibly to predict their behavior.

The main challenge faced in terms of modeling by differential equations is to formulate the equations describing the problem from a set of limited information about the general behavior of the system. However, since a possible solution to the proposed modeling problem (equation 2) is available, it is possible to analyze it, based on some assumptions, and determine the differential equations that represent the language model problem.

Considering that the overall goal of the interpolation model is to integrate the benefits of each language model and assuming the language modeling as a natural system problem, it is possible to treat this problem as the solution of a partial differential equation, i. e.,

$$\frac{\partial u(x,y)}{\partial x} + \frac{\partial u(x,y)}{\partial y} = u(x,y), \quad (10)$$

where $u(x,y)$ is the interpolation model and x and y are the independent variables, representing the Word-based n-gram and POS-based language models, respectively.

Analyzing (2) it is possible to see that this equation is not a particular solution of (10), but a particular solution of another partial differential equation, namely

$$\begin{cases} x\dfrac{\partial u(x,y)}{\partial x} + y\dfrac{\partial u(x,y)}{\partial y} = u(x,y), \\ u(1,1) = 1. \end{cases} \quad (11)$$

To solve (11), it is possible to use the technique of separation of variables, which reduces the partial differential equation to several ordinary differential equations [29]. In this case, it is assumed that a solution can be expressed as the product of two unknown functions, where each one is only function of the respective independent variable. This assumption seems very reasonable, after a study considering each language model as an independent problem. Thus, it is possible to have

$$u(x,y) = X(x)Y(y) \Rightarrow \begin{cases} \dfrac{\partial u(x,y)}{\partial x} = X'Y \\ \dfrac{\partial u(x,y)}{\partial y} = XY' \end{cases} \quad (12)$$

and

$$x\frac{\partial u(x,y)}{\partial x} + y\frac{\partial u(x,y)}{\partial y} = xX'Y + yXY' = XY. \quad (13)$$

Dividing (13) by XY, it follows that

$$x\frac{X'}{X} + y\frac{Y'}{Y} = 1. \quad (14)$$

Since X is only a function of the variable x and Y of y, the two terms in (14) should be constant, and equal to α and

[1]For a more detailed explanation see [28].

$(1-\alpha)$ (known as *separation constant*[29]), or

$$x\frac{X'}{X} = 1 - y\frac{Y'}{Y} = \alpha, \tag{15}$$

which implies in two ordinary equations, namely

$$x\frac{X'}{X} = \alpha \Rightarrow X(x) = \gamma_1 \cdot x^{\alpha} \tag{16}$$

and

$$1 - y\frac{Y'}{Y} = \alpha \Rightarrow Y(y) = \gamma_2 \cdot y^{(1-\alpha)}, \tag{17}$$

where γ_1 e γ_2 are two constants.

Finally, substituting (16) and (17) in (12), one obtains

$$u(x,y) = \gamma_1 \cdot x^{\alpha} \cdot \gamma_2 \cdot y^{(1-\alpha)} = \gamma \cdot x^{\alpha} y^{(1-\alpha)}, \tag{18}$$

where $\gamma = \gamma_1 \cdot \gamma_2$ is also a constant.

Applying the condition $u(1,1) = 1$ to (18), it follows that

$$u(1,1) = \gamma = 1, \tag{19}$$

and the solution will be

$$u(x,y) = x^{\alpha} \cdot y^{(1-\alpha)}, \tag{20}$$

which is the same as the geometric interpolation presented in equation 2.

In the same context, considering that the word n-gram model (x in the differential partial modeling) plays a fundamental role in predictive modeling systems and can be improved in combination with a POS-based language model, it is here proposed a modified partial differential equation, given by

$$\begin{cases} x\dfrac{\partial u(x,y)}{\partial x} + \dfrac{\partial u(x,y)}{\partial y} = u(x,y), \\ u(1,1) = 1. \end{cases} \tag{21}$$

This assumption, that word n-gram model is regarded as the main model, can be verified by testing our Word Prediction System when the model are swapped, i.e. considering the n-gram model as model y and the POS model as model x, as it can be seen in section IV.

As originally proposed by [30], which assumes an exponential form for the statistical distribution in language models, the modeling performed in equation 21 seeks to create a natural exponential function to interpolate the word n-gram model and the POS-based language model. In contrast to the works presented in [30], [31], which also seeks to construct a stochastic model that represents the behavior of the random process using the concept of maximum entropy model, the differential partial modeling presented in equation 21, common in the modeling of natural processes as radioactive decay and population decay (both have as a solution a natural exponential response, usual in natural processes) [29], can provide different analyzes of the possible solutions, telling us how *physical parameters*, *initial* and *boundary conditions* could affect these solutions. Besides, the mathematical formulation proposed allows us to find the solution at a single point (x,y) without going through the entire marching process of finding the solution at all other points.

In a similar way to the geometric interpolation modeling, the new interpolation model can be found by solving (21), whose solution is

$$u(x,y) = \gamma \cdot x^{\alpha} \cdot e^{(1-\alpha)y}. \tag{22}$$

Again, applying the condition $u(1,1) = 1$ to (22), one has

$$u(1,1) = 1 = \gamma \cdot e^{(1-\alpha)} \Rightarrow \gamma = \frac{1}{e^{(1-\alpha)}}, \tag{23}$$

and, finally,

$$u(x,y) = \frac{1}{e^{(1-\alpha)}} \cdot x^{\alpha} \cdot e^{(1-\alpha)y}. \tag{24}$$

Rewriting (24) in terms of the language models, it follows that

$$\begin{aligned} P^{(proposed)}_{interpolation}(w_i|w_{i-(n-1)}, t_{i-(m-1)}) = \\ \frac{1}{e^{(1-\alpha)}} \cdot [P_{n-gram}(w_i|w_{i-(n-1)})]^{\alpha} \cdot \\ \cdot e^{(1-\alpha)\cdot[P_{m-POS}(w_i|w_{i-(m-1)})]}, \end{aligned} \tag{25}$$

where α can be empirically obtained.

It is worth noticing that (25) has a form that is similar to the conventional Maximum Entropy model first developed in [14], with the major difference in using the n-gram model as $P_{n-gram}(w_i|w_{i-(n-1)})^{\alpha}$. In such work, the authors confront two essential tasks of statistical modeling: to determine a set of statistics that captures the behavior of a random process, and to combine these facts into an accurate model of the process – a model capable of predicting the future process output. Trying to solve this problem, the proposed methodology, based on differential equations, is quite interesting, since it addresses the problem of building interpolation models and opens the way for the use of different mathematical tools to analyze the language modeling process. Besides, as expected in natural processes, our model also has the natural exponential response with a similar *idea* to the Verhulst equation [32], who confronted the idea that the human population tends to grow in a geometric progression, proposing a (still somewhat arbitrary) partial differential equation whose solution is also based in a natural exponential response.

IV. Experiments

The tests and subsequent analyzes needed to confirm the assumptions made are presented in this section. Firstly, to construct possible comparisons between the methodology adopted here with the Linear and Geometric combination models, it was necessary to select appropriate training, validation and test sets, besides using proper procedures for testing each methodology.

The text set used for training and testing is one of the key aspects in the evaluation step, since it may influence significantly the results. Thus, texts from newspapers and texts transcript from spoken language were chosen to compose the test set. Texts from newspapers were adopted because they have a language dedicated to a great number of readers, having a reasonable contextual diversity in terms of vocabulary and grammatical constructions. Texts transcript from spoken language, by their turn, were adopted because they are more

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2547743, IEEE/ACM Transactions on Audio, Speech, and Language Processing

spontaneous, less rigid, and cover the daily communication in general.

The evaluation procedures for the word prediction were conducted by automatic methods, where all language models were incorporated into a WP system and experiments, common in these types of systems.

It is also important to consider that any change in the configuration parameters of the experiment (language, test and training texts, WP interface system) can lead to significant variations in the results. This variability makes it very difficult to compare the results presented here with others already established (in [19] the main factors that can affect the prediction results on a given system are exposed and discussed).

A. Training Set

The training sets used to train the language models and to generate the dictionaries for each language here addressed (Portuguese, Spanish and English) are shown in Table II.

TABLE II
NUMBER OF WORDS USED IN THE TRAINING SET FOR EACH LANGUAGE MODEL.

Language	**Language Model**		***Corpus***
	n-gram	m-POS	
Portuguese	17,599,914	505,412	CHAVE [33]
Spanish	17,601,472	502,800	Spanish Gigaword First Edition [34]
English	17,763,503	511,066	English Gigaword [35]

B. Validation Set

The validation set used to found the best coefficients α to combine the language models (n-gram and m-POS) are shown in Table III, with Tables V, VI and VII showing the values of KSS and PP for English, Portuguese and Spanish language, respectively, with α ranging between 0.0 to 1.0 for all the interpolation models. Furthermore, in order to validate our proposed methodology of combining POS-based models we perform a test using the validation set of Table III and the results are presented in Table IV for all languages and compared with the traditional POS-based model.

TABLE III
NUMBER OF WORDS USED IN THE VALIDATION SET TO FOUND THE BEST α VALUES.

Language	**#Words**	**#Keystroke Needed**	***Corpus***
Portuguese	80,259	498,310	Rhetalho [36]
Spanish	67,722	409,088	Corpus92 [37]
English	81,631	515,197	MASC [38]

As can be seen in Table IV, the proposed m-POS language model showed improvements in the KSS and PP parameters when compared to the traditional m-POS language model. However, the proposed m-POS language model performs worst in all parameters when compared to the results obtained by the n-gram model, as can be seen in the Tables V, VI and VII. This may be related to the size of the training set used by each model, i.e., for the n-gram model it was used a very large training set with more than 17 million words against only 500 thousand words used to train the POS-based model for each language. These results reinforce the idea of using the n-gram as the main models of our interpolation model and opens the way to search for new class-based models as the recurrent neural network proposed by [39].

TABLE IV
KSS AND PP RESULTS TO THE PROPOSED m-POS INTERPOLATION MODEL COMPARED WITH THE TRADITIONAL POS-BASED MODEL USING 1 AND 5 WORDS IN THE PREDICTION LIST. BOLD NUMBERS INDICATE THE BEST RESULTS.

Language	**m-POS Language Model**	**5 words**		**1 word**	
		KSS(%)	PP	KSS(%)	PP
English	Traditional	38.01	6901.0	26.50	6193.7
	Proposed	**38.12**	**4357.9**	**26.70**	**3919.4**
Portuguese	Traditional	48.62	6711.8	35.98	5978.6
	Proposed	**48.65**	**4429.4**	**36.08**	**3969.9**
Spanish	Traditional	45.52	6728.2	32.47	5528.9
	Proposed	**45.62**	**4444.6**	**32.57**	**3661.6**

TABLE V
VALIDATION SET VARYING α FOR EACH INTERPOLATION MODEL CONSIDERING 1 AND 5 WORDS IN THE ENGLISH PREDICTION LIST. BOLD NUMBERS INDICATE THE BEST RESULTS COMPARED WITH THE BASELINE.

Language Model	**5 words**			**1 word**		
	α	PP	KSS(%)	α	PP	KSS(%)
n-gram	-	370.1	41.48	-	152.2	30.48
Linear	0.0	1732.1	38.58	0.0	1658.3	26.53
	0.1	683.9	40.40	0.1	354.9	28.98
	0.2	559.90	40.91	0.2	275.8	29.80
	0.3	499.10	41.18	0.3	235.0	30.18
	0.4	456.90	41.34	0.4	209.0	30.41
	0.5	425.30	41.45	0.5	190.2	30.57
	0.6	400.30	41.53	0.6	175.1	30.64
	0.7	379.20	41.57	0.7	163.8	30.74
	0.8	361.40	41.61	0.8	154.1	30.77
	0.9	**344.50**	**41.61**	**0.9**	**145.3**	**30.78**
	1.0	370.1	41.48	1.0	152.2	30.48
Log-linear	0.0	1732.1	38.58	0.0	1658.3	26.53
	0.1	1509.9	38.73	0.1	954.0	26.60
	0.2	1264.6	39.12	0.2	780.2	26.67
	0.3	1060.3	39.55	0.3	638.2	27.41
	0.4	893.6	40.00	0.4	520.1	27.95
	0.5	754.6	40.40	0.5	422.7	28.55
	0.6	640.8	40.78	0.6	341.6	29.16
	0.7	545.7	41.12	0.7	274.8	29.74
	0.8	464.7	41.38	0.8	220.9	30.28
	0.9	**394.0**	**41.55**	**0.9**	**176.0**	**30.63**
	1.0	370.1	41.48	1.0	152.2	30.48
Proposed	0.0	27.9	39.17	0.0	27.3	28.06
	0.1	187.3	41.05	0.1	143.9	30.61
	0.2	199.6	41.15	0.2	143.4	30.69
	0.3	213.1	41.25	0.3	143.3	30.74
	0.4	228.2	41.35	0.4	143.1	30.79
	0.5	244.4	41.43	0.5	142.7	30.80
	0.6	261.9	41.53	0.6	142.2	30.82
	0.7	279.8	41.59	0.7	141.4	30.82
	0.8	298.0	41.64	0.8	140.8	30.85
	0.9	**314.9**	**41.65**	**0.9**	**139.7**	**30.85**
	1.0	370.1	41.48	1.0	152.2	30.48

It is possible to note in Tables V, VI and VII that our proposed language model starts with perplexity values lower than the baseline to all languages, however, these values do not provide the best results for KSS. Even seemingly counter-

TABLE VI
VALIDATION SET VARYING α FOR EACH INTERPOLATION MODEL CONSIDERING 1 AND 5 WORDS IN THE PORTUGUESE PREDICTION LIST. BOLD NUMBERS INDICATE THE BEST RESULTS COMPARED WITH THE BASELINE.

Language Model	5 words			1 word		
	α	PP	KSS(%)	α	PP	KSS(%)
n-gram	-	373.2	53.56	-	169.2	40.98
Linear	0.0	1346.2	50.97	0.0	1297.6	37.21
	0.1	567.7	53.02	0.1	301.5	40.37
	0.2	475.9	53.56	0.2	245.8	41.17
	0.3	430.6	53.86	0.3	216.6	41.62
	0.4	398.5	54.02	0.4	197.0	41.86
	0.5	374.4	54.14	0.5	182.5	42.09
	0.6	353.8	54.20	0.6	171.0	42.23
	0.7	336.9	54.26	0.7	162.0	42.35
	0.8	322.3	54.30	0.8	153.8	42.39
	0.9	**308.5**	**54.30**	**0.9**	**146.9**	**42.40**
	1.0	373.2	53.56	1.0	169.2	40.98
Log-linear	0.0	1346.2	50.97	0.0	1297.6	37.21
	0.1	1141.5	51.41	0.1	699.7	38.08
	0.2	973.0	51.80	0.2	589.6	38.47
	0.3	834.0	52.22	0.3	497.3	38.89
	0.4	715.4	52.68	0.4	418.3	39.41
	0.5	617.8	53.09	0.5	350.9	39.98
	0.6	534.5	53.46	0.6	294.4	40.59
	0.7	465.0	53.82	0.7	246.2	41.19
	0.8	403.6	54.08	0.8	205.7	41.75
	0.9	**348.4**	**54.25**	**0.9**	**171.3**	**42.21**
	1.0	373.2	53.56	1.0	169.2	40.98
Proposed	0.0	24.0	51.30	0.0	22.2	39.21
	0.1	187.6	53.48	0.1	141.4	42.06
	0.2	197.6	53.64	0.2	141.9	42.17
	0.3	208.9	53.79	0.3	142.2	42.26
	0.4	221.2	53.93	0.4	142.4	42.30
	0.5	234.4	54.08	0.5	142.8	42.38
	0.6	247.6	54.17	0.6	143.0	42.42
	0.7	260.9	54.24	0.7	142.7	42.44
	0.8	274.4	54.29	0.8	142.4	42.46
	0.9	**286.6**	**54.34**	**0.9**	**141.8**	**42.47**
	1.0	373.2	53.56	1.0	169.2	40.98

TABLE VII
VALIDATION SET VARYING α FOR EACH INTERPOLATION MODEL CONSIDERING 1 AND 5 WORDS IN THE SPANISH PREDICTION LIST. BOLD NUMBERS INDICATE THE BEST RESULTS COMPARED WITH THE BASELINE.

Language Model	5 words			1 word		
	α	PP	KSS(%)	α	PP	KSS(%)
n-gram	-	292.2	49.63	-	111.8	36.70
Linear	0.0	1590.4	46.35	0.0	1495.4	32.15
	0.1	531.2	48.62	0.1	259.5	35.55
	0.2	435.6	49.29	0.2	203.8	36.35
	0.3	386.4	49.60	0.3	174.6	36.80
	0.4	354.8	49.78	0.4	155.0	37.04
	0.5	330.4	49.93	0.5	140.5	37.27
	0.6	308.1	50.02	0.6	129.0	37.41
	0.7	291.3	50.09	0.7	120.0	37.53
	0.8	276.2	50.11	0.8	111.8	37.57
	0.9	**265.7**	**50.12**	**0.9**	**105.3**	**37.58**
	1.0	292.2	49.63	1.0	111.8	36.70
Log-linear	0.0	1590.4	46.35	0.0	1495.4	32.15
	0.1	1103.5	47.20	0.1	657.5	33.24
	0.2	935.0	47.59	0.2	547.4	33.63
	0.3	796.0	48.01	0.3	455.1	34.05
	0.4	677.4	48.47	0.4	376.1	34.57
	0.5	579.8	48.88	0.5	308.7	35.14
	0.6	496.5	49.25	0.6	252.2	35.75
	0.7	427.0	49.61	0.7	204.0	36.35
	0.8	365.60	49.88	0.8	163.5	36.91
	0.9	**308.50**	**50.06**	**0.9**	**129.1**	**37.37**
	1.0	292.2	49.63	1.0	111.8	36.70
Proposed	0.0	28.8	47.53	0.0	27.7	34.68
	0.1	144.7	49.53	0.1	102.1	37.35
	0.2	154.7	49.69	0.2	102.6	37.46
	0.3	166.0	49.84	0.3	102.9	37.46
	0.4	178.3	49.98	0.4	103.1	37.59
	0.5	191.5	50.13	0.5	103.5	37.67
	0.6	204.7	50.22	0.6	103.7	37.71
	0.7	218.8	50.27	**0.7**	**103.6**	**37.71**
	0.8	**230.6**	**50.28**	0.8	102.2	37.70
	0.9	242.2	50.22	0.9	104.1	37.56
	1.0	292.2	49.63	1.0	111.8	36.70

intuitive, this behavior has already been found in other studies like [40] and [41] to speech recognition systems. In [40] and [42] different language models were constructed and their output was combined using interpolation weights chosen to satisfy a maximum likelihood criterion (and hence to minimize perplexity). This led to models which had considerably lower perplexities than the baseline trigram model, but no decrease in Word Error Rate (WER).

In order to better understand the counterintuitive behavior of our methodology, we can use an example trying to take a closer look to the contribution of α, the n-gram and the POS-based language models in the score assigned by each combination model in our word prediction system. Figure 3 shows a sample example extracted from our validation set considering $\alpha = 0.0$ and $\alpha = 0.9$. It is important to mention that in this example the correct word to be predicted is "*what*".

It can be seen in Figure 3 considering $\alpha = 0.0$ that since our proposed model uses a natural exponential approach, even a low probability value associated to the POS-based model generates a high score value, which does not occur for the linear and geometric models. Besides, as α increases ($\alpha = 0.9$), it is also possible to note in Figure 3 a decrease in the probability of the correct word, going from 0.3692 to 0.1681 in our interpolation model, but, with an improvement in the rank position, moving from third to second position in the word prediction list. This kind of improvement in the position also occurs to the others interpolation models also showing an improvement in the probability of the correct word. These results are in agreement with the work presented by [43], which shows that the word probability has a strong correlation to its rank and less correlation to the entropy of the distribution in the test set.

We can also associate this counterintuitive behavior in the fact that not having used a normalization factor in equation 25. As in [44], using the model as part of a classifier (e.g., a speech recognizer or a word prediction system) does not require knowledge of this normalization factor, because the relative ranking of the different classes is not changed by a single, universal constant. However, as described by [45], as a result, we will no longer have *probabilities* in our model but instead *scores*, and we can no longer calculate perplexities. Nonetheless, in order to compute perplexity, the same authors proposes in [44] a *perplexity reduction ratio* to estimate reduction in per-word perplexity over the baseline. In this way, we also calculate the PP based on the perplexity reduction ratio proposed by [44] and the results are shown in Table VIII. It is important to mention that the methodology proposed by [44] does not change the KSS values found by our proposed language model.

Fig. 3. Sample example of the prediction system to English with five words in the prediction list and considering $\alpha = 0.0$ and 0.9.

TABLE VIII
PP RESULTS BASED ON THE PERPLEXITY REDUCTION RATIO PROPOSED BY [44] APPLIED TO OUR EXPONENTIAL LANGUAGE MODEL, VARYING α AND CONSIDERING 1 AND 5 WORDS IN THE ENGLISH, PORTUGUESE AND SPANISH PREDICTION LISTS. BOLD NUMBERS INDICATE THE BEST RESULTS COMPARED WITH THE BASELINE.

Language	**5 words**			**1 word**		
	α	PP	KSS	α	PP	KSS
English	n-gram	370.1	41.48	n-gram	152.2	30.48
	0.0	353.4	39.17	0.0	151.4	28.06
	0.1	343.1	41.05	0.1	147.8	30.61
	0.2	335.7	41.15	0.2	147.0	30.69
	0.3	329.7	41.25	0.3	145.4	30.74
	0.4	322.7	41.35	0.4	144.1	30.79
	0.5	319.8	41.43	0.5	141.8	30.80
	0.6	316.4	41.53	0.6	139.6	30.82
	0.7	315.0	41.59	0.7	138.2	30.82
	0.8	313.5	41.64	0.8	138.0	30.85
	0.9	**311.3**	**41.65**	**0.9**	**137.5**	**30.85**
	1.0	370.1	41.48	1.0	152.2	30.48
Portuguese	n-gram	373.2	53.56	n-gram	169.2	40.98
	0.0	355.2	51.30	0.0	168.4	39.21
	0.1	341.5	53.48	0.1	165.5	42.06
	0.2	333.2	53.64	0.2	162.8	42.17
	0.3	324.2	53.79	0.3	159.6	42.26
	0.4	319.2	53.93	0.4	155.0	42.30
	0.5	318.0	54.08	0.5	148.8	42.38
	0.6	313.5	54.17	0.6	145.0	42.42
	0.7	312.1	54.24	0.7	144.9	42.44
	0.8	310.6	54.29	0.8	144.0	42.46
	0.9	**308.1**	**54.34**	**0.9**	**143.7**	**42.47**
	1.0	373.2	53.56	1.0	169.2	40.98
Spanish	n-gram	292.2	49.63	n-gram	111.8	36.70
	0.0	275.3	47.53	0.0	111.2	34.68
	0.1	265.0	49.53	0.1	108.5	37.35
	0.2	255.1	49.69	0.2	105.2	37.46
	0.3	251.9	49.84	0.3	103.6	37.46
	0.4	250.4	49.98	0.4	102.5	37.59
	0.5	247.5	50.13	0.5	99.9	37.67
	0.6	245.1	50.22	0.6	96.7	37.71
	0.7	243.1	50.27	**0.7**	**96.4**	**37.71**
	0.8	**242.2**	**50.28**	0.8	97.5	37.70
	0.9	245.7	50.22	0.9	102.6	37.56
	1.0	292.2	49.63	1.0	111.8	36.70

As can be seen, the results in Table VIII are more intuitive than those shown in Tables V, VI and VII, since the PP values decreases as increases the contribution of the n-gram model. Thus, we also calculate the PP in the test set based in the method proposed by [44] with the results shown in section IV-F.

We can also see that most experiment groups in Tables V, VI, VII and VIII achieve their best results when α is set to 0.9 and in general, the performance monotonically increases with $alpha$. Again, this results may be related to the size of the training set used by each model, i.e., for the n-gram model it was used a very large training set with more than 17 million words against only 500 thousand words used to train the POS-based model.

Finally, in the brief validation test to follow, we tried to verify how the overall system will perform if the models are swapped, i.e., considering the n-gram model to be model y and POS-based model to be model x in equation 24. The main idea here is to emphasize the importance of the n-gram model to the proposed interpolation model validating the mathematical formulation that led to the general language model based on partial differential equations. Thus, the results are showed in Table IX for English, Portuguese and Spanish using the validation set showed in Table III, with five words in the prediction list and considering $\alpha = 0.0$, 0.5 and 1.0.

As expected, since the m-POS language model showed worse results than the n-gram language model (Table IV), the "swapped" model also showed worse results than the n-gram model. Besides, it is possible to note in Table IX that the KSS for all language decreases as α increases. Again, these results reinforce the idea of using the n-gram as the main models of our interpolation model.

C. Test Set

Here, it is important to use texts that were not used in the training and validation sets. This way, for Portuguese some texts from the journalistic *corpus* TeMário [46] and the

TABLE IX
KSS AND PP RESULTS WITH 5 WORDS IN THE PREDICTION LIST CONSIDERING THE n-GRAM MODEL TO BE MODEL y AND POS-BASED MODEL TO BE MODEL x IN EQUATION 24. BOLD NUMBERS INDICATE THE BEST RESULTS.

Language	Language Model	α	PP	KSS(%)
English	n-gram	-	**370.1**	**41.48**
	Proposed	0.0	23.1	39.15
		0.5	316.7	38.27
		1.0	4357.9	38.12
Portuguese	n-gram	-	**373.2**	**53.56**
	Proposed	0.0	20.0	50.37
		0.5	297.9	49.38
		1.0	4429.4	48.65
Spanish	n-gram	-	**292.2**	**49.63**
	Proposed	0.0	23.3	47.14
		0.5	320.4	46.29
		1.0	4444.6	45.62

"Português Falado" *corpus* [47] were chosen, which contains texts transcript from audio recordings of the language spoken in Brazil. To compose the test set for Spanish, the Europarl *corpus* [48] was used, which contains texts transcript from speeches of the European Parliament, and texts from the HC *corpus* [49], consisting of newspaper articles from different sources. Finally, for English the test set was extracted from the Brown *corpus* [50] and the Uppsala Student English *corpus* [51], consisting of newspapers and transcript texts, respectively. Table X shows the topic, number of words and keystroke needed (without word prediction) to write the texts in each test set.

TABLE X
TOPIC AND NUMBER OF WORDS USED IN THE TEST SET.

Language	#Words	#Keystroke Needed	*Corpus*
Portuguese	72,222	434,581	TeMário [46] Português Falado [47]
Spanish	102,981	637,895	Europarl [48] HC *corpus* [49]
English	90,931	520,923	Brown [50] Uppsala Student English [51]

It is important to mention that about 3% of the words in each test set are out-of-vocabulary (OOV). For the moment special attention was dedicated to the known words and not about unknown words. This is treated as a separated problem and smoothing techniques were used to avoid null probabilities for the unseen events in the test set.

D. Word Prediction Engine

In order to evaluate our method, the software PREDWIN was used [19]. This software was firstly developed for Spanish and was adapted here for Portuguese and English. This system has some important blocks, such as:

- **user model:** it is the automatic algorithm used by the word prediction system to emulate a real user. For each letter in the test text the prediction system shows a list of predicted words. If the desired word is in this list, the user model selects the word. If not, it selects the next letter. This loop is repeated until the test text ends;
- **coordination module:** it controls the flow of information between the user interface/user model and the dictionaries and prediction methods. It obtains the word prediction list that each method provides and sends the most adequate to the user interface;
- **dictionaries:** contain the words and the information about each word required to support all the word prediction methods, such as POS tags and word frequencies.

It is also important to highlight some system parameters that will affect the word prediction:

- **coefficient α:** it can be defined as a variable to combine language models. This variable can take values between 0 and 1, giving more weight to any one of the language models. As shown in [17] and [18], this parameter has been experimentally determined by varying its value from 0.1 to 0.9 with increments of 0.1;
- **back-off regression:** the language models also has certain limitations when the frequency of word or POS sequences are null. To overcome such limitations, some techniques can be used as the back-off technique. Initially proposed in [52], when the language model has zero frequency it is approximated by the immediately previous model and continues until reaching the unigram model. Due to its simplicity of implementation and satisfactory results, as presented in [19] to the Spanish language, this methodology is also used in this work;
- **number of previously used POS and words:** the number POS and words (in n-gram model) previously used has a major influence on the system, since these are directly related to the structure of the language. However, as the basic statistical models, it is necessary a reasonable number of training data to ensure reliable models. The works presented in [17], [19] and [53], which has a m-POS and n-gram training set with size similar to that used in this paper, obtained the best results using up to two previous POS. Thus, the models in this work were evaluated using two previous categories and the *back-off regression* previously explained;
- **number of words in the prediction list:** according to [19], 7 is the maximum number of words that the user can process, maintaining the balance between increased cognitive load and the increasing processing time in the selection of words predicted. However, other studies, such as [54] and [55], showed that the optimal number of words presented to the user must be at most 5 words. In this work the prediction systems were tested using 1 and 5 words in the prediction list. With just 1 word in the prediction list it is possible to simulate an automatic system, common in mobile devices, for example;
- **number of words in the dictionaries:** the general dictionaries play a fundamental role in the word prediction system. Once categorized texts are needed, the same texts used for training the syntactic models presented in Table II are considered. Table XI shows, for each language, the

number of words in general dictionaries;

TABLE XI
NUMBER OF WORDS FOR EACH GENERAL DICTIONARY.

Language	#Words
Portuguese	118,878
Spanish	92,482
English	57,711

- **repeated suggestions in consecutive predictions:** since we are working with an ideal user, its possible to drop the predicted words not previously selected by the user. According to [53], this methodology enables an increase in the probability to suggest the most appropriate words;
- **automatic insertion of blank spaces:** apart from word prediction, there are certain characteristics that can be automatically inserted to improve the KSS. An example of this technique is the insertion of blank spaces after punctuation symbols (comma, period and colon, for example);
- **test of significance:** this work attempts to validate the methods using a statistical test based on the calculation of confidence intervals for proportions, given by

$$p = \sigma \cdot \sqrt{\frac{KSS_{LM} \cdot (1 - KSS_{LM})}{N}}, \qquad (26)$$

where p is the confidence interval, KSS_{LM} is the keystroke saved by each language model, N is the total number of keystroke needed to write the text without word prediction and σ is a constant parameter that depends on the confidence interval, usually set to 1.96 (or 95% of confidence). Thus, the KSS in the experiment is within the range

$$KSS_{LM} = KSS_{LM} \pm p, \qquad (27)$$

and it will be considered significantly better (with respect to the baseline) if the results are better and, moreover, the confidence intervals do not overlap.

E. Performance Measures

The WP system was evaluated according to four different criteria: keystroke saved (KSS), hit rate (HR), words predicted (WP) and perplexity (PP).

To better illustrate some performance measures, Figure 4 shows an example of the Word Prediction System for English with five words in the prediction list. In this case the system searches for the most probable words, based on the probability given by the proposed interpolation model, that starts with letter *(*d). Once we know what the next word in the test text is, the number of predicted words (WP) can be easily accounted by analyzing the possible words given in the prediction list. In Figure 4, for example, we know that the next word in the test text is "*day*", which is also in the list of predicted words making it to have a correctly predicted word count.

The KSS is referred to the percentage of keystrokes that the user saves when using the word prediction system. It is calculated by comparing two measures: the total number of keystrokes needed to type the text (K_T) without the help of the word prediction and the effective number of keystrokes needed when using word prediction (K_E). Hence,

$$KSS = \frac{K_T - K_E}{K_T} \times 100. \qquad (28)$$

The higher the value of KSS is, better is the system performance.

Fig. 4. Sample example of the prediction system to English with five words in the prediction list.

The HR is defined as the percentage of correct words that appear in the suggestion list before entering any letter of the following word. In other words, it is the relation between the number of times that a word is guessed without knowing any letter and the total number of words in the test text. Again, a higher HR means a better performance.

The PP can be measured with the cross entropy calculated on a test set with N words, given by

$$H(\mathbf{W}) = \frac{1}{N} \sum_{i=1}^{N} \log \left(\frac{1}{P\left(w_i | w_{i-(n-1)}, t_{i-(m-1)}\right)} \right), \qquad (29)$$

where $\mathbf{W} = (w_1, w_2, \ldots, w_N)$ represents the words in the test set, and the PP is

$$PP(\mathbf{W}) = 2^{H(\mathbf{W})}. \qquad (30)$$

In other words, the PP can be defined as the average number of potential choices/words after a given string of words [17].

To better understand the difference between the calculation of the PP measure with one or five words in the prediction list, suppose the example showed in Figure 4. Here the correct word was predicted with a probability of 0.70 and in the first position on the prediction list. Assuming now that the correct word is "*deal*", the probability for this word is 0.07 and it is in the third position on the prediction list. Thus, the main difference relies in the fact that with only one word in the prediction list the WP system always proposes the most probable word. In contrast, with five words in the prediction list, the system may propose words with low probabilities. This also explains the fact that the total PP in WP systems with only one word in the prediction list is lower than WP systems with more words in the prediction system.

F. Results

In order to evaluate the WP system with different interpolation models, the experiments have been conducted with the texts shown in Table X. The results are presented in Table XII, considering the word n-gram model as baseline. The relative improvements were also evaluated and the results are presented in Table XIII, along with the test of significance for each model. Table XIV shows the PP results using the perplexity reduction ratio proposed by [44].

G. Discussion

It is possible to notice in Table XII that all interpolation models show improvements with respect to the number of KSS compared to the word-based n-gram model. These results are in agreement with [56] and supports that the problem of word prediction can be improved by finding linguistically relevant factors and an efficient method is the combination of a POS-based and word-based language models. In some cases, especially when the results obtained by language models with 5 suggested words are considered, the influence of the syntactic model is clear, i.e., even with the decrease in the number of words predicted, there is an improvement in KSS.

It is also important to notice in Table XII that the proposed interpolation model shows the best results in all parameters related to word prediction (WP, HR and KSS) considering 1 word in the prediction list. Here, it is important to mention that when considering the absolute HR values for Portuguese language the proposed method achieves 16,692 keystroke saved against 16,680 saved by the liner method, 16,675 saved by the geometric method and 16,665 saved by the baseline method. When considering 5 words in the prediction list, the word n-gram models present, in all languages, the best results for the WP parameter. These results clearly show the importance of word n-gram models in the prediction of *function words*, consisting mainly of pronouns, determiners, preposition and auxiliary verbs, and words with a lower set of letters, as compared to the *content words*, normally nouns, verbs, adjectives and adverbs [57].

When analyzed with the PP values in Table XII, considering 5 suggested words, the proposed interpolation model also shows the best results reaching 20.69%, 25.92% and 13.85% relative reduction for the Spanish, Portuguese and English. With 1 word in the prediction list, the proposed model reaches 7.21%, 17.79% and 7.76% relative reduction in the PP. This measure is consistent for almost all models, i.e., it is also possible to notice in Table XII that the PP obtained by the geometric model for the English and Spanish languages is higher than the value obtained by the word n-gram model, even with higher values of KSS. Such inconsistencies have been already presented and discussed in other works, such as [17]. In the same context, analyzing Table XIV the perplexity reduction ratio proposed by [44] reaches relative reduction of 11.82%, 16.21% and 9.65%, with 1 word, and 18.11%, 19.45% and 16.25%, with 5 words, for the Spanish, Portuguese and English, respectively.

In Table XIII it is possible to notice that the relative improvements in English using the proposed interpolation model (1.66% and 0.57% for 1 and 5 words in the prediction list) are lower than those obtained for the other languages. These results make clear the statement that English is not as *morphologically rich* than Portuguese and Spanish languages, and are consistent with the work in [53]. Using a linear combination (with $\alpha = 0.8$) to combine a word bigram model and a POS-based model considering two previous POS, trained using 81 million words and 5.8 million categorized words, respectively, [53] presents a total of 53.14% KSS against 52.90% KSS obtained by the word bigram model itself, considering 5 words in the prediction list, a test set of $951,932$ words and a POS tagset with 79 classes.

It is also worth to analyze in Table XIII the relative improvements obtained by each interpolation model considering the number of words in the prediction list. In this case, it can be seen that increasing the number of suggested words reduces the relative improvement of each model, but increases the total number of KSS. It is also possible to notice that some of the results for English, excluding the results obtained by the proposed model, were not significant. A possible solution to this could be to increase the number of words in the test set.

The empirically optimized values for the coefficient α in Tables V, VI and VII also deserve discussion. It can be observed that the values founded for all interpolation models were almost equal to 0.9 (the proposed model showed a alpha value of 0.8 and 0.7 in Table VII for Spanish with 5 and 1 suggested words, respectively). These results are consistent with [17] and [21] and opens the way to search for optimization methods to determine the separation coefficient α, as the Expectation Maximization (EM) proposed in [58].

V. Conclusion

In an attempt to improve the word prediction task, we have proposed a natural exponential interpolation model, which combines a traditional word-based n-gram language model with a POS-based language model, defined as the linear combination of three different POS-based languages (with each weight coefficient based on the AUC). We address this problem by first finding a partial differential equation to represent the language modeling, which will be used to derive the interpolation model.

The proposed methodology was evaluated on three different languages: Portuguese, Spanish and English and the results reported in this paper show improvements in the KSS, HR and PP parameters, with 1 and 5 words in the prediction lists, and show the benefits of integrating various sources of information under the partial differential equation framework for improving language modeling. When analyzing the PP values, the perplexity reduction ratio proposed by [44] seems more intuitive since the PP values decreases as increases the contribution of the n-gram model.

While this paper focused on combining a word n-gram and a m-POS based language model, it is worth noting that there is a growing body of work using continuous-space models in a variety of language processing tasks, particularly for deriving semantic representations of words as described in [59] and more recently as in [60]. Then, future efforts can

TABLE XII
WORD PREDICTION RESULTS TO THE DIFFERENT INTERPOLATION MODELS WITH 1 AND 5 WORDS IN THE PREDICTION LIST.

		#Words in the Prediction List							
		1				**5**			
Language	**Model**	HR(%)	WP(%)	KSS(%)	PP	HR(%)	WP(%)	KSS(%)	PP
Spanish	n-gram	21.30	86.93	41.32	95.6	38.00	**95.37**	54.09	240.2
	Linear	21.20	87.43	42.30	90.1	38.10	95.22	54.64	218.4
	Geometric	21.20	87.30	42.13	111.1	38.00	95.21	54.58	253.0
	Proposed	**21.40**	**87.66**	**42.44**	**88.7**	**38.30**	95.28	**54.76**	**190.5**
Portuguese	n-gram	23.30	86.16	38.29	164.1	**35.61**	**95.17**	50.35	416.3
	Linear	23.30	86.43	39.27	139.2	35.50	94.90	51.14	331.6
	Geometric	23.30	86.40	39.11	162.7	35.50	94.89	51.08	373.3
	Proposed	**23.30**	**86.74**	**39.37**	**134.9**	35.60	94.93	**51.19**	**308.4**
English	n-gram	21.10	88.17	38.88	121.2	35.10	**97.24**	52.01	283.7
	Linear	21.10	88.50	39.46	115.8	35.10	97.17	52.28	263.6
	Geometric	21.00	88.06	39.18	141.0	35.10	97.22	52.21	304.7
	Proposed	**21.11**	**88.63**	**39.52**	**111.8**	**35.20**	97.20	**52.31**	**244.4**

TABLE XIII
TEST OF SIGNIFICANCE AND IMPROVEMENTS RELATIVE TO THE n-GRAM MODEL FOR RESULTS SHOWN IN TABLE XII.

		#Words in the Prediction List					
		1			**5**		
Language	**Model**	KSS(%)	Relative Improvement(%)	Significant	KSS(%)	Relative Improvement(%)	Significant
Spanish	n-gram	41,32($\pm$0.12)	-	-	54,09($\pm$0.12)	-	-
	Linear	42,30($\pm$0.12)	2.39	Yes	54,64($\pm$0.12)	1.02	Yes
	Geometric	42,13($\pm$0.12)	1.98	Yes	54,58($\pm$0.12)	0.91	Yes
	Proposed	**42,44**($\pm$0.12)	**2.72**	Yes	**54,76**($\pm$0.12)	**1.24**	Yes
Portuguese	n-gram	38,29($\pm$0.15)	-	-	50,35($\pm$0.15)	-	-
	Linear	39,27($\pm$0.15)	2.58	Yes	51,14($\pm$0.15)	1.57	Yes
	Geometric	39,11($\pm$0.15)	2.16	Yes	51,08($\pm$0.15)	1.45	Yes
	Proposed	**39,37**($\pm$0.15)	**2.83**	Yes	**51,19**($\pm$0.15)	**1.67**	Yes
English	n-gram	38,88($\pm$0.13)	-	-	52,01($\pm$0.14)	-	-
	Linear	39,46($\pm$0.13)	1.49	Yes	52,28($\pm$0.14)	0.52	No
	Geometric	39,18($\pm$0.13)	0.79	Yes	52,21($\pm$0.14)	0.37	No
	Proposed	**39,52**($\pm$0.13)	**1.66**	Yes	**52,31**($\pm$0.14)	**0.57**	Yes

TABLE XIV
PP RESULTS BASED ON THE PERPLEXITY REDUCTION RATIO PROPOSED BY [44] APPLIED TO OUR EXPONENTIAL LANGUAGE MODEL IN THE TEST SET.

Language	**5 words**	**1 word**
	PP	PP
English	237.6	109.5
Portuguese	335.3	137.5
Spanish	196.7	84.3

be concentrated on improving the proposed partial differential equation by adding more information such as the semantic-based model, searching for a more sophisticated interpolation for the language models.

It is also interesting to carefully study the impact of the parameter α on the proposed interpolation model and present evaluations on different domain. Finally, we plan to test our current interpolation language model on another tasks, such as automatic language recognition.

REFERENCES

[1] R. Foulds, "Communication rates of non-speech expression as a function in manual tasks and linguistic constraints." in *In Proceedings of the International Conference on Rehabilitation Engineering*. Toronto: RESNA, 1980, pp. 83–87.

[2] N. Garay-Vitoria and J. Abascal, "Text prediction systems: a survey," *Univers. Access Inf. Soc.*, vol. 4, no. 3, pp. 188–203, Feb. 2006.

[3] M. Ghayoomi and S. Momtazi, "An overview on the existing language models for prediction systems as writing assistant tools," in *Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on*, San Antonio, Texas, 11-14 October 2009, pp. 5083–5087, iSSN: 1062-922X.

[4] P. Vyrynen, "Perspectives on the utility of linguistic knowledge in english word prediction," Ph.D. dissertation, University of Oulu, Linnanmaa, November 19th 2005.

[5] H. AlMubaid, "A learning-classification based approach for word prediction," *Int. Arab J. Inf. Technol.*, vol. 4, no. 3, pp. 264–271, 2007.

[6] N. Garay-Vitoria and J. Abascal, "Modelling text prediction systems in low- and high-inflected languages," *Comput. Speech Lang.*, vol. 24, no. 2, pp. 117–135, 2010.

[7] J. L. Arnott and N. Alm, "Towards the improvement of augmentative and alternative communication through the modelling of conversation," *Computer Speech and Language*, vol. 27, no. 6, pp. 1194–1211, 2013, special Issue on Speech and Language Processing for Assistive Technology.

[8] K. Hacioglu and W. Ward, "On combining language models: oracle approach," in *Proceedings of the first international conference on Human language technology research*, ser. HLT '01. Stroudsburg, PA, USA: Association for Computational Linguistics, 2001, pp. 1–4. [Online]. Available: http://dx.doi.org/10.3115/1072133.1072210

[9] D. Linares, J.-M. Benedí, and J.-A. Sánchez, "A hybrid language model based on a combination of n-grams and stochastic context-free grammars," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 3, no. 2, pp. 113–127, 2004.

[10] T. Brychcn and M. Konopk, "Semantic spaces for improving language modeling," *Computer Speech and Language*, no. 0, 2013, in Press. doi:

10.1016/j.csl.2013.05.001.

[11] M. A. Chachoo and S. M. K. Quadri, "Adaptive hybrid pos cache based semantic language model," *International Journal of Computer Applications*, vol. 39, no. 13, pp. 7–10, February 2012, published by Foundation of Computer Science, New York, USA.

[12] C.-H. Chueh, J.-T. Chien, and H.-M. Wang, "A maximum entropy approach for semantic language modeling," *Computational Linguistics and Chinese Language Processing*, vol. 11, no. 1, pp. 37–56, March 2006.

[13] S. Wang, D. Schuurmans, and Y. Zhao, "The latent maximum entropy principle," *ACM Trans. Knowl. Discov. Data*, vol. 6, no. 2, pp. 8:1–8:42, Jul. 2012.

[14] A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra, "A maximum entropy approach to natural language processing," *Comput. Linguist.*, vol. 22, no. 1, pp. 39–71, Mar. 1996. [Online]. Available: http://dl.acm.org/citation.cfm?id=234285.234289

[15] S. Wang, S. Wang, R. Greiner, D. Schuurmans, and L. Cheng, "Exploiting syntactic, semantic and lexical regularities in language modeling via directed markov random fields," in *In Proceedings of ICML 2005*, 2005, pp. 948–955.

[16] E. Cambria and B. White, "Jumping nlp curves: A review of natural language processing research," *IEEE Comp. Int. Mag.*, vol. 9, no. 2, pp. 48–57, 2014.

[17] K. Trnka, "Word prediction techniques for user adaptation and sparse data mitigation," Ph.D. dissertation, University of Delaware, Newark, DE, USA, 2011, iSBN: 978-1-124-48009-1.

[18] A. Fazly and G. Hirst, "Testing the efficacy of part-of-speech information in word completion," in *TextEntry '03: Proceedings of the 2003 EACL Workshop on Language Modeling for Text Entry Methods*. Budapest, Hungary: Association for Computational Linguistics, 2003, pp. 9–16.

[19] S. E. Palazuelos-Cagigas, "Contribution to word prediction in spanish and its integration in technical aids for people with physical disabilities," Ph.D. dissertation, Universidad de Alcalá de Henares, Alcalá de Henares, Madrid, Spain, 2001.

[20] J. R. Bellegarda, "Exploiting Latent Semantic Information in Statistical Language Modeling," *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1279–1296, Aug. 2000.

[21] T. Wandmacher and J.-Y. Antoine, "Methods to integrate a language model with semantic information for a word prediction component," *CoRR*, vol. abs/0801.4716, 2008.

[22] A. Carlo, C. Nicola, Mancarella, Paolo, and R. Michele, "A word predictor for inflected languages: system design and user-centric interface," in *Proceedings of the Second IASTED International Conference on Human Computer Interaction*, ser. IASTED-HCI '07. Anaheim, CA, USA: ACTA Press, 2007, pp. 148–153.

[23] D. C. Cavalieri, T. F. Bastos-Filho, M. Sarcinelli-Filho, S. E. Palazuelos-Cagigas, J. M. Guarasa, and J. L. M. Sánchez, "A part-of-speech tag clustering for a word prediction system in portuguese language," *Procesamiento del Lenguaje Natural*, vol. 47, pp. 197–205, 2011, iSSN: 1135-5948.

[24] E. Bick, "The parsing system "palavras": Automatic grammatical analysis of portuguese in a constraint grammar framework," Ph.D. dissertation, Aarhus University, Aarhus, Denmark, November 2000.

[25] ——, "A constraint grammar parser for spanish," in *Proceedings of the International Joint Conference IBERAMIA/SBIA/SBRN 2006 - 4th Workshop in Information and Human Language Technology (TIL 2006)*, Ribeiro Preto, October 27-28 2006, iSBN: 85-87837-11-7.

[26] A. Voutilainen and J. Heikkila, "An English constraint grammar (EngCG): a surface-syntactic parser of English," in *Creating and Using English Language Corpora*, U. Fries, G. Tottie, and P. Schneider, Eds., vol. 13. Amsterdam: Editions Rodopi, 1994.

[27] S. E. Palazuelos-Cagigas, J. L. Martn-Snchez, J. Macas-guarasa, J. C. Garca-Garca, D. C. Cavalieri, T. F. Bastos-Filho, and M. Sarcinelli-Filho, "Machine learning methods for word prediction in brazilian portuguese," in *Everyday Technology for Independence and Care*, ser. Assistive Technology Research Series. IOS Press, 2011, vol. 29, ch. Chapter 4: Advanced Technologies, pp. 424 – 431.

[28] T. Fawcett, "An introduction to roc analysis," *Pattern Recogn. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.

[29] W. Boyce and R. DiPrima, *Elementary Differential Equations*. Wiley, 2012.

[30] S. D. Pietra, V. J. D. Pietra, and J. D. Lafferty, "Inducing features of random fields," *CoRR*, vol. 1, 1995.

[31] S. D. Pietra, V. J. D. Pietra, J. Gillet, J. D. Lafferty, H. Printz, and L. Ures, "Inference and estimation of a long-range trigram model." in *ICGI*, ser. Lecture Notes in Computer Science, R. C. Carrasco and J. Oncina, Eds., vol. 862. Springer, 1994, pp. 78–92.

[32] N. Bacar, "Verhulst and the logistic equation (1838)," in *A Short History of Mathematical Population Dynamics*. Springer London, 2011, pp. 35–39.

[33] D. Santos and P. Rocha, "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*, Bath, UK, September 15-17 2004, pp. 821-832.

[34] D. Graff, "Spanish gigaword first edition," Linguistic Data Consortium, Philadelphia, 2006.

[35] D. Graff and C. Cieri, "English gigaword," Linguistic Data Consortium, Philadelphia, 2003.

[36] N. Neto, C. Patrick, A. Klautau, and I. Trancoso, "Free tools and resources for brazilian portuguese speech recognition." *J. Braz. Comp. Soc.*, vol. 17, no. 1, pp. 53–68, 2011.

[37] U. P. F. I. U. de Lingstica Aplicada (IULA), "Corpus92 corpus," http://hdl.handle.net/10230/20054, 2012, accessed on 05/07/2014.

[38] N. Ide, C. Fellbaum, C. Baker, and R. Passonneau, "The manually annotated sub-corpus: A community resource for and by the people," in *Proceedings of the ACL 2010 Conference Short Papers*, ser. ACLShort '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 68–73.

[39] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, "Recurrent neural network based language model," in *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, 2010, pp. 1045–1048.

[40] P. Clarkson and A. J. Robinson, "Language model adaptation using mixtures and an exponentially decaying cache," in *In Proceedings of ICASSP-97*, 1997, pp. 799–802.

[41] P. Clarkson and T. Robinson, "The applicability of adaptive language modelling for the broadcast news task," in *ICSLP98*, 1998, pp. 233–236.

[42] ——, "Towards improved language model evaluation measures," in *In Proceedings of EUROSPEECH 99, 6th European Conference on Speech Communication and Technology*, 1999.

[43] N. A. Smith, "Adversarial evaluation for models of natural language," *CoRR*, vol. abs/1207.0245, 2012.

[44] R. Rosenfeld, S. F. Chen, and X. Zhu, "Whole-sentence exponential language models: a vehicle for linguistic-statistical integration." *Computer Speech & Language*, vol. 15, no. 1, pp. 55–73, 2001.

[45] S. Chen, K. Seymore, and R. Rosenfeld, "Topic adaptation for language modeling using unnormalized exponential models," in *Proceedings of ICASSP '98*, 1998.

[46] T. A. S. Pardo and L. H. M. Rino, "TeMrio: Um corpus para sumarizao automtica de textos," Ncleo Interinstitucional de Lingstica Computacional (NILC), So Carlos-SP, Srie de Relatrios do NILC NILC-TR-03-09, Outubro 2003, 11 p.

[47] M. F. B. do Nascimento, L. Pereira, and J. Saramago, "Portuguese corpora at clul," in *Second International Conference on Language Resources and Evaluation*, Atenas, Grécia, 2000, pp. 1603-1607.

[48] P. Koehn, "Europarl: A Parallel Corpus for Statistical Machine Translation," in *Conference Proceedings: the tenth Machine Translation Summit*, AAMT. Phuket, Thailand: AAMT, 2005, pp. 79–86.

[49] H. Christensen, "Hc corpora," http://corpora.heliohost.org/, 2012.

[50] W. N. Francis and H. Kucera, "Brown corpus manual," Department of Linguistics, Brown University, Providence, Rhode Island, US, Tech. Rep., 1979.

[51] M. W. Axelsson, "Project use (uppsala student english)," ASLA Information, pp. 25–26, 1999.

[52] S. M. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer," in *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1987, pp. 400–401.

[53] A. Fazly, "The use of syntax in word completion utilities," Master's thesis, University of Toronto, Department of Computer Science, 2002.

[54] D. K. Anson, P. Moist, M. Przywara, H. Wells, H. Saylor, and H. Maxime, "The effects of word completion and word prediction on typing rates using on-screen keyboards," *Proceedings RESNA'05*, 2005.

[55] K. Trnka, D. Yarrington, J. McCaw, K. F. McCoy, and C. Pennington, "The effects of word prediction on communication rate for aac," in *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, ser. NAACL-Short '07. Stroudsburg, PA, USA: Association for Computational Linguistics, 2007, pp. 173–176.

[56] C. Wilson, "Combining part of speech induction and morphological induction," Ph.D. dissertation, University of Melbourne, Melbourne, Australia, Novembro 2004.

[57] V. den Bosch, "Scalable classification-based word prediction and confusible correction," *Traitement Automatique des Langues*, vol. 46, no. 2, pp. 39–63, 2006.
[58] D. Jurafsky and J. H. Martin, *Speech and Language Processing (2nd Edition)*, 2nd ed. Prentice Hall, Maio 2008.
[59] J. A. Botha and P. Blunsom, "Compositional morphology for word representations and language modelling," *CoRR*, vol. abs/1405.4273, 2014.
[60] H. Fang, M. Ostendorf, P. Baumann, and J. B. Pierrehumbert, "Exponential language modeling using morphological features and multitask learning." *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 12, pp. 2410–2421, 2015.

Daniel Cruz Cavalieri received the B.Sc. degree in Electrical Engineering from the Federal University of Viçosa, Viçosa, MG, Brazil, in 2005, and the Ph.D. in Electrical Engineering from the Federal University of Espírito Santo, Vitória, ES, Brazil, in 2013. He is with the Department of Control and Automation Engineering, Instituto Federal do Espírito Santo. His research interests are Natural Language Processing, Human-Machine Interfaces and Assistive Technologies.

Sira E. Palazuelos-Cagigas is an Associate Professor at the University of Alcala, Madrid, Spain. She got her Ph.D. in Telecommunication Engineering (2001) and the B.Sc. degree in Telecommunication Engineering (1994) from the Technical University of Madrid. She has participated in over 50 research projects funded by public and private organizations, including 14 as principal investigator. Has published 12 papers in international and national journals, and has presented more than 80 papers in international and national conferences. She also owns 2 utility models, and 3 patents (currently under examination). She has directed a PhD thesis and has been part of the organizing committee of several national and international conferences or special sessions organized by the Department of Electronics of the University of Alcala. She has also co-authored several international book chapters. Throughout her career she has performed 5 research stays, including a 6-month stay at the International Computer Science Institute (ICSI), Berkeley, USA in 2003.

Teodiano Freire Bastos-Filho graduated in Electrical Engineering (Universidade Federal do Espírito Santo, Vitória, Brazil) in 1987, and received the Ph.D. degree in Physical Sciences (Universidad Complutense de Madrid, Spain) in 1994. He is with the Department of Electrical Engineering, Universidade Federal do Espírito Santo, and with the Brazilian National Council for Scientific and Technological Development (CNPq). His research interests are signal processing, rehabilitation robotics and assistive technologies.

Mário Sarcinelli-Filho received the B.S. degree in Electrical Engineering from Federal University of Espírito Santo, Brazil, in 1979, and the M. Sc. and Ph. D. degrees, also in Electrical Engineering, from Federal University of Rio de Janeiro, Brazil, in 1983 and 1990, respectively. He is currently an Associate Professor at the Department of Electrical Engineering, Federal University of Espírito Santo, Brazil, and a researcher of the Brazilian National Council for Scientific and Technological Development (CNPq). His research interests are signal and image processing, sensing systems, mobile robot navigation, coordinated control of mobile robots and unmanned aerial vehicles. He has co-authored 37 journal papers, 266 conference papers, and 7 book chapters. He has also advised 15 PhD students and 18 MSc students.