GRAMMATICAL AND STATISTICAL WORD PREDICTION SYSTEM FOR SPANISH INTEGRATED IN AN AID FOR PEOPLE WITH DISABILITIES

Palazuelos, S., Aguilera S., Rodrigo J., Godino J.
Laboratorio de Tecnologías de Rehabilitación. ETSI de Telecomunicación, Universidad Politécnica de Madrid. 28040-Madrid, Spain
{sira, aguilera, rodrigo, godino}@die.upm.es

ABSTRACT

In this paper, we describe and evaluate recent work and results achieved in the word prediction system for Spanish introduced in [3] and [4], applying both statistical and grammatical methods: word pairs and trios, bipos and tripos models, and a stochastic context free grammar. The predictor is included in several software applications, to enhance the typing rate for users with motorical problems, who can only use a switch for writing. These users have difficulties in writing, and usually in communicating with other people, and the inclusion of word prediction in the system allow them to increase their typing rate from 2-3 words per minute up to 8-10 words per minute.

1. WORD PREDICTION IN AIDS FOR PEOPLE WITH DISABILITIES

Generally speaking, a word prediction application tries to find out which is the word a user is typing or is going to type, before he/she writes it completely. The guessed words are shown somehow, so that, if the desired one is included in that list, the user can select and insert it in the text, avoiding the need to type the rest of the word, thus reducing the time and effort necessary to write the text. This is especially useful for slow typists, or people who are not able to use a conventional keyboard. The methods to avoid its use usually involve the utilization of as many switches as the user is able to handle (usually 1 or 2) and a keyboard emulator, controlled with the switches. Because of the little versatility of the switch-based access, the use of matrices scanning is generally employed. As this is a very slow input method, several acceleration techniques are used to increase the user's typing rate. Word prediction is one of them.

People with other writing problems may also take advantages of the word prediction: dyslexic users or people with frequent spelling errors feel more confident writing with the help of this aid. Children and people learning a second language are also target users of these systems.

Several applications based on word prediction have been described in the literature, each one devoted to a particular group of users, with their specific needs, as we can see in [3][4][5][6][7]. Some of them are speech prostheses to help non-vocal users to communicate: tools including speech synthesis in which user enters text, and it is spoken. Word prediction increases the communication rate of these users, especially if they have also motorical problems that make them write slowly. Other tools assist users when entering texts in a computer. In this paper we will evaluate the effect of the prediction in this last group of applications.

2. WORD PREDICTION METHODS

Several prediction methods have been implemented, using different information sources. All of them are used together in the system, in order to make use of the advantages of each one. In this section we include a description of the main components of the word prediction system.

2.1 General static information sources

The basic component is a general dictionary: it is a 40000-word database containing statistical and grammatical information. It is the basis for the generation of all the probabilities involved in the prediction process, in both the grammatical and statistical methods.

Each entry in the dictionary consists of a word, its part-of-speech, and the number of times this particular pair appeared in the training corpus (the number of times that word appeared with that tag in the corpus). The training corpus was manually tagged by an expert, assigning each word in its context one of 18 selected categories, and its features. For example, each noun also contains information of its gender and number, which is very important in the prediction using grammatical information in Spanish, because it allows filtering of nouns that do not agree with the previous words, improving the word prediction quality.

If a word may be labeled with different parts of speech, depending on the context, it will appear several times in this dictionary, each time with the appropriate frequency. This information is essential in the grammatical methods, allowing us to consider all the possibilities (until we have enough information for the disambiguation).

From the training corpus, bipos and tripos matrices were automatically generated. This basic information was the first grammatical prediction method implemented in the system, allowing the prediction of the next word's part-of-speech from the information of the previous words. After finding out the most likely part-of-speech to follow the previous words, the most frequent words belonging to those categories were predicted.

Both the general dictionary and the bipos and tripos matrices are static knowledge sources, and they cannot be trained or customized by the user.

2.2 User adaptive prediction methods

As well as these static methods, user information is essential in word prediction. Knowing information about the topic, or, if possible, the whole text the user is writing, may help in several ways: on the one hand, previous words are usually repeated in the texts, especially the ones specific of the topic the user is writing about. On the other hand, it is very difficult to obtain a complete general dictionary, including every word in the language and, even in that case, the frequencies of the words, or its preferred part of speech may change within each subject. These are the main reasons to include the possibility of adaptation.

Two different customization methods have been implemented. The first one is a learning module, which dynamically generates a dictionary with the text the user is writing, at the same time he/she is typing. It records the new words (for example, specific words of the subject the user is writing about), and updates the probabilities associated to each word, giving priority to the last ones used in the text (recency). After some sentences from the beginning of the session, the vocabulary offered to the user is better adapted to the topic of interest. One of the main advantages of this method is the ability to learn new words: once one word is typed, it will be presented to the user, even prevailing over the ones in the general dictionary.

At the same time, word pairs are recorded, (for very frequent function words, trios of words are also stored). This information is used when the first word of the pair/trio is written again: then, the system shows the words that appeared after it in the text (the second word in the corresponding word pair/trio), before the user writes any other letter. It is very frequent that any of these words is the right one, and can be selected directly, thus saving effort. After that, the system shows the words that followed this last word, starting the process again. It is very common that small groups of words appear together in the text several times, that is the reason why the adaptation is a very efficient prediction method. (For example, in this paper: "prediction methods", "very efficient", "in the text", etc.).

The second customization method is based on the possibility of generating customized dictionaries, using texts previously written about them (by the user or not). For example, in case the user wants to write about football of politics, he/she may generate a dictionary from newspaper texts about them. Then, the new vocabulary and frequencies will be updated [8] before the user starts typing, increasing his/her speed. Word pairs and trios are also stored.

These two adaptive methods work with individual dictionaries, apart from the general one. The general dictionary is not modified, to avoid the possibility of recording spelling mistakes, (very frequent when using scanning based interfaces, which are common applications of the word prediction).

2.3 Grammatical word prediction

The prediction methods presented until now use only statistical information (general dictionary, and learning modules), and simple grammatical information (bipos and tripos). With them, the prediction of each word is only based on information about the two previous words. The next step is to predict taking into account information of the whole sentence which is being written. This may be accomplished by including a linguistic module which is able to model long term relationships between words.

The inclusion of a complex language model has additional advantages for certain users with problems to generate grammatically correct sentences, (for example, for people with dyslexia, with linguistic disabilities, or for students learning a second language), because they know that words predicted by the system are grammatically correct. As well as this, the quality of the word prediction increases, and user's opinion about the quality of the word prediction is better if impossible/ungrammatical predictions are removed from the prediction list.

The language module implemented is based on a stochastic context free grammar (SCFG) for Spanish, which is currently being developed. Although a SCFG grammar does not cover all the correct sentences, it is powerful enough to describe the most common structures of Spanish (or any natural language), and it is restricted enough so that efficient parsers can be built to analyze sentences.

The mechanism chosen to handle the grammar is a statistical chart parser [1], which was modified to behave as a part-of-speech predictor. The main advantage of using a chart parser is its efficiency in comparison with other parsing techniques.

The behavior of the parser is bottom up when analyzing the user's text, because it processes the user's input, checking if it matches the rules of the language, until the last full word written by the user. The prediction process is top down, because it checks for the active rules (the rules whose first part matches the previous words), and predicts that the category of the next word is the part-of-speech expected for that rules.

In the parsing/prediction process, there are many different factors to take into account: several rules to expand a symbol, ambiguity of words, errors, etc. We handle these factors by adding frequency information to the whole process, and choosing the most probable options. In the following sections we will explain how each factor affects the probability of every rule and predicted word.

Word prediction process

Word prediction is made by calculating the probability of each word in the dictionary to be the following *(Prob(W))*, and choosing the most probable words. When using grammatical knowledge, we calculate the probability of each category to be the following, and then we predict words belonging to the preferred parts of speech using *(Prob(W/C)*Prob(C))*.

Prob(W/C) is the *lexical probability:* the probability of predicting W knowing that the next category is C. It is extracted from the corpora/dictionaries, as:

$$\Pr(W/Ci) = \frac{\mathrm{Freq}\,(W \cap Ci)}{\mathrm{Freq}\,(Ci)}$$

With Freq(X)= number of times X appears in the text. We use Ci instead of C, because each word may be tagged with different parts-of-speech in the dictionary. In the prediction process, we take this into account by changing the calculations to add the probability of the word to belong to each category:

$$\text{Prob (W)} = \sum (\text{Prob}(W/C_i)*\text{Prob}(C_i))$$

The estimation of *Prob(C_i)* depends on the grammatical method used. In our case, is extracted from the parser, being *Prob(C_i)* the probability of all the branches which need C_i to be extended. Several factors influence these probabilities:

Many rules may expand the same non-terminal item, and not all of them have the same frequency in the language, so we have assigned each rule its probability to expand the non-terminal symbol.

$$\text{Prob(Ri)} = \frac{Freq(Ri)}{Freq(Symbol)}$$

So, the probability of each branch is the product of the probabilities of all the rules involved in it. With this, and the possibility that several rules expect the same category in a particular moment, we compound the following expression to calculate the probability of the next part of speech:

$$\Pr ob_1(C_i) = \sum_{\substack{\text{Branches}\\ \text{expecting Ci}}} \left(\prod_{\substack{\text{Rules in}\\ \text{the branch}}} \text{Prob}(R_i) \right)$$

After the user types a word, the corresponding rules are expanded, according to that word, but we have to consider again that a word may belong to several categories. In each context, only one of the possible parts of speech is valid, so we need an ambiguity resolution process. In this parser, our approach is to keep all the categories that extend any of the branches, avoiding the selection of any of them until we have more information, (more words are typed). As the probability of the word to belong to each category is different (*Prob(C_i/W)*), we will take this into account by multiplying the probability of each rule by *Prob(C_i/W)*, being C_i the category expected by it.

$$\Pr(C_i/W) = \frac{\text{Freq}(W \cap C_i)}{\text{Freq}(W)}$$

So, with each new word, the probability of each branch is updated to:

$$\Pr(C_i) = \sum_{\substack{\text{Branches}\\ \text{expecting } C_i}} \left(\prod_{\substack{\text{Rules in}\\ \text{the branch}}} \Pr(R_i) \prod_{\substack{\text{Words which}\\ \text{have extended } R_i}} \Pr(C_i/W_j) \right)$$

This method will allow keeping all the categories, giving preference to the rules followed by the most probable categories of the words. The selection of one of the categories (ambiguity resolution) is made by the following words, which expand certain rules, and not others, eliminating, probably, the rules extended by non-appropriate categories. Ideally, at the end of the sentence, only a branch of the parser will be active, and each word will be assigned the right part of speech in that context. However usually there are several possible trees at the end. In that case, the one with higher probability is chosen.

Error recovery strategies

In the previous sections we have explained the usual operation mode of the parser, when the sentence agrees the grammar. Unfortunately, parsing errors are so common that can degrade the parser efficiency significantly if error recovery strategies are not implemented. There are mainly two reasons that makes the parsing process fail: either it does not know the word it is analyzing, or the word does not match any active rule. Each error is handled in a different way.

Unknown words will extend every rule in the parser, without changing its probability, as substitution errors in [2]. In "unknown words" we include spelling errors and correct words not included in the dictionary, because we have no technique to distinguish them. In following versions, we will change this to avoid expanding rules that expect words belonging to closed categories (when all the words belonging to that part of speech are listed in the dictionary). But we should be careful with it, because of the spelling errors.

The lack of coverage of the rules is solved by backing off to the bipos and tripos models previously seen, until the user writes a word with a category that can start a sentence, and then restarts the parsing process. In parallel with the case of the words, this lack of rules coverage may be because not all the rules are listed in our grammar, or because the sentence is not grammatically correct.

2.4 Additional techniques

As well as the word prediction methods previously seen, there are another two techniques also included in the system, which may also influence the results. One is the suffixes prediction: as well as words, suffixes are also predicted (this is especially important when writing words not included in the dictionary).

The other technique is a filter that stores the predictions presented to the user, that have been rejected, so they are not shown again while writing the current word.. This filter is especially efficient with long words, when there are several very similar words, differing only in the final part (i.e. correspondent, correspondence). The user may want to write one of them, and the other appears in the menu from the first letters, and it does not disappear when adding more letters because the beginning of the word is the same. After one or two times the word is presented to the user, the system may assume he/she has seen the word, and it is not the correct, so it will not be presented again (although it matches also the following letters the user will type). With this, some positions are freed in the menu, accelerating the presentation of the right word.

3. RESULTS

The evaluation of word prediction systems as writing aids considers its efficiency, measured in two different ways, depending on its application. If it is included in aids for people with motorical problems, (to reduce the effort needed to write), we are interested in the savings in the number of keystrokes needed to write the text. If it is included in aids for people with linguistic problems, we evaluate the "prediction coverage", which can be defined as the number (or percentage) of the words that are predicted (a word is considered as *predicted* if it is proposed by the system before the user finishes typing it). Further refining this measurement, we may even calculate the number of letters typed until the right word is shown to the user. Systems with better results in word coverage may be worst in keystroke savings, because they may very good at predicting short words (that saves few keystrokes).

We have performed several tests, exactly in the same conditions, changing only the prediction method used, to evaluate the results of each particular technique, and several combinations, find out which is the best one.

To obtain the results, we have established the parameters that may influence the prediction performance measurements: the number of keystrokes saved/typed are counted as in a normal keyboard (one keystroke per small letter, two per capital letter, etc. making it independent of the user interface). The maximum number of predicted words has been set to 7, the maximum number of suffixes to 4, and the number of times a word is shown before rejecting it to 1. A 25673 words test text has been used.

In these conditions, the results obtained using an automatic evaluation module are (in these tests we are have not used the SCFG, as the number of rules defined so far is too small to be useful):

Prediction method	#Keystrokes (% savings)	Prediction coverage	
		Words (%)	Suffixes (%)
Without prediction	175777	0 (0%)	0 (0%)
Gram (no adaptive)	105084 (40,21%)	19468 (75.83%)	1991 (7.76%)
Adaptive (no Gramm)	85023 (51.63%)	22515 (87.70%)	791 (3.08%)
Gram. + adaptive	84396 (51.98%)	22451 (87.45%)	802 (3.12%)

With: *Gram* = Use of grammatical word prediction.
Adaptive = Learning module active.
Words coverage: the words correctly predicted
Suffixes coverage: the words completed with the prediction of suffixes.

As we can see, adaptive methods are responsible for most of the improvement achieved, when compared with grammatical ones. This is because of the close adaptation to the vocabulary, expressions, and frequencies in the test text, which is big enough (25000 words) for a nearly perfect adaptation. This is also the reason for which there is so little difference between using or not grammatical information along with adaptive methods. In order to properly evaluate the actual advantages of both approaches in a real application, smaller test texts should be used. The idea is that users will probably never write such long texts in the same session, and grammatical information is especially useful at the beginning of the writing session, when the system does not know the user's style of writing, and only general information can be applied. In those cases, grammatical methods show their better behavior.

4. CONCLUSIONS AND FUTURE WORK

In this paper we have seen a brief description of the prediction methods used in a grammatical and statistical word prediction system for Spanish. We have presented in more detail the parser used to increase the grammatical knowledge used in the prediction process. An evaluation of the prediction performance has been presented, showing that adaptive methods produce better results than grammatical models in isolation, but the best technique consists of the combination of both methods, getting up to 52% in keystroke savings, and a 90% in words + suffixes coverage.

Lot of work is to be done on developing and refining the set of rules for the SCFG. Testing with bigger test texts will also be addressed, to get more significant results. Regarding the word prediction methods, we are planning the use of new techniques, such as neural networks for word class prediction, and improving the features of the chart parser.

5. ACKNOWLEDGEMENTS

This work has been supported by Comunidad Autónoma de Madrid, Ref.: CAM 07T/0010/1997

6. REFERENCES

1. Allen, J, "Natural Language Understanding", 1995, Second Ed. Benjamin/Cummings, CA.
2. Kong Joo Lee et all. "A Robust Parser Based on Syntactic Information". Proceedings of 7th Conference of the EACL, Dublin, Ireland, Pgs.: 223-228, 1995.
3. Palazuelos, S.E., Montero, J.M., Gómez S., Aguilera S., "On the Development of a Word Processor With Word Prediction for Severely Physically Handicapped, Non-Vocal Users: PREDICE". Proceedings of ECART 3, Pgs:119-121, Lisbon. October 1995.
4. Palazuelos, S. E., Godino, J. I., Aguilera, S., "Comparison between adaptive and non adaptive word prediction methods in a word processor for motorically handicapped non vocal users", Proceeding of the AAATE 1997, Pgs.: 158 – 162 Greece.
5. Tyvand, S., Demasco, P., "Syntax statistics in word prediction", Proc. ECART2, Stockholm, 26-28 May 1993, pp 11.2
6. Hunnicutt, S., "Input and output alternatives in word prediction" STL-QPSR 2-3/1987, pp 15-29
7. Claypool T, Booth L, Ricketts IW and Gregor P. "BT ASSIST - A tool to help you with your reading, writing and arithmetic", ISAAC 1996, Vancouver.
8. Palazuelos, S., Aguilera, S., "Report on Word Prediction for Spanish",. VAESS project Deliverable WP7T3D.2IR .TIDE N. 1174. 1996