Evaluation of prediction methods applied to an inflected language

Nestor Garay-Vitoria, Julio Abascal-González
Laboratory of Human-Computer Interaction for Special Needs
University of the Basque Country
Manuel Lardizabal 1; E-20018 Donostia
E-mails: {nestor, julio}@si.ehu.es
Phone: +34 943 018000; Fax: +34 943 219306

Keywords: *Prediction, Language inflection, Morphemes, Keystroke savings, Hit ratio*

Abstract
Prediction is one of the techniques that have been applied to Augmentative and Alternative Communication to help people enhancing the quality and quantity of the composed text in a time unit. Most of the literature has been focused in word prediction methods that may be easily applied to non-inflected languages. However, for inflected languages other approaches that mainly distinguish roots and suffixes may enhance the results (in terms of keystroke savings and hit ratio) of predictive systems. In this paper we present the approaches we have applied to the Basque language (an inflected one) and the results they achieve with a particular text (that was not used while creating the initial lexicons the systems use for prediction). Starting from this evaluation, one of the presented approaches is suggested as the best one.

1.- Introduction
People with severe motor and oral disabilities that use Augmentative and Alternative Communication systems in order to write the messages they want to make known need a high time to compose them. The communication speed they achieve is very low (as some authors have estimated it is near to 2-10 words/minute) in comparison to the speed achieved in spoken conversations (estimated in 120-180 words/minute) [Alm-92].
This situation may be alleviated if the required time to compose messages is reduced. One of the techniques to do this is the anticipation or prediction, which consists in trying to guess what a person is going to tell or write immediately after what it has been told or written [Garay-94]. In this way, apart from enhancing the communication speed, the physical effort required to compose messages may be reduced because with fewer keystrokes the same information quantity can be composed.
Even the study of prediction techniques was initially devoted to alleviate some of the needs of people with disabilities, people without disabilities may take benefit of them if they are adequately integrated in applications and have an appropriate interface [Darragh-90][1].
However, there is not an unique valid approach to all languages. Most of the references in the literature are related to the case of non-inflected languages (like English, French or Spanish), where roots have a small amount of variations (for example, in the number: spy/spies). In this case, as the number of different forms of a root is small, it is possible to include all of them in the dictionary the system requires to work. In these languages, the block that is normally treated by the system is the word (they are word prediction methods). Unfortunately, in inflected languages (like Basque, German or Swedish) roots have a huge amount of variations and it is nearly impossible to store all of them in the dictionary[2]. So the prediction of a word is made in some steps; typically, roots and suffixes

---

[1] *These applications try to minimise the user's effort using the computer. For example, enhancing the quality and/or the quantity of the composed text, like it may be the spelling correction.*

[2] *For example, in the case of the Basque language a noun may have 62 basic declensions, but as it is an agglutinative language, suffixes may be recursively concatenated. Then, with two levels of recursion it has been estimated that a noun may have 458,683 forms [Agirre-92].*

are distinguished and then a word is completed in a two-step process (one step for the root and the other for the suffix). A complete revision of the problem may be found in [Garay-97, Garay-01]. Then, in inflected languages the prediction treats morphemes and they are morpheme prediction methods.
Next, we are going to present the methods that we have studied in order to apply to the Basque language (an inflected one). Their main characteristics are expressed and the results they achieve are presented and compared, trying to determine which one is the best.

2.- Prediction techniques studied for the Basque language

In this section we are going to briefly present the prediction methods that we have applied in the case of the Basque. They have been fully presented in [Garay-01].

2.1.- Prediction by frequencies

The first methods that we have studied are based in purely statistical information in order to determine which are the blocks the systems offer to the user as proposals.

A.- Word prediction

As a starting point we consider the word as the block to be treated by the predictor and the system has a lexicon where the words and their frequencies are stored. This system offers the most frequent words that start with the string the user has written so far, in the same way they are stored in the lexicon. This operational way can be used in any language, if a lexicon of the required characteristics is included in the system. All the required adaptations are automatically made updating the frequencies.
The main objectives of this first approach are to establish the involved knowledge in the problem environment and to make comparisons with the rest of the approaches.

B.- Morpheme prediction

In this approach, there are two lexicons: one for roots and the other for the suffixes, each one containing the frequency apparitions of the associated morphemes. From this approach on, both the system and the user have to clearly distinguish roots and suffixes. One way to clearly distinguish if the system is working with roots in including in the interface some dots at the end of the lemmas (for example, “know…”); in the same way, the system will be working with suffixes if the proposals are preceded by some dots (for instance, “…ing”). The cognitive interface has also to offer the feature of explicitly changing from root prediction to suffix prediction. This change is implicitly made when the proposals are accepted or when the user ends composing the treated unit (root or suffix), maybe with a final space or a similar character.
This system starts proposing the most frequent roots that start with the string the user has written so far until that root is guessed or entirely written; after that, it proposes the suffixes until guessed or completed the word. Then, it starts again with the roots and so on. Here too, updating the frequencies makes the adaptations.
Anyway, as the case of the verbs is special in the Basque (their most probable suffixes are very different in relation to those of the nouns), new approaches that take into account syntactic information are tried.

2.2.- Morpheme prediction using probability tables taking into account categories

Still distinguishing roots and suffixes, some syntactical information is added to the lexicons. Moreover, there are some bidimensional tables that store the relative probabilities of the categories of the suffixes to appear after the categories of the roots and vice versa for making predictions.
In this system the most probable roots starting with the written string whose categories are the most probable after the last composed suffix are proposed; after being composed the root, the suffixes are treated, and so on. Updating the frequencies of the lexicons and the entries of the used tables makes the adaptation.
However, there are some problems to solve, as there are ambiguous morphemes and new roots and suffixes may appear if adapted to user's vocabulary. Some possible ways to solve

them may be found in the literature [Garay-01]: some systems typically try to guess the category, others directly ask the missed information to the user or some give a special category (related to the new roots or suffixes) in order to be completed the information when the user is ready to do that (maybe by some special user or in a special session).

2.3.- Prediction using syntactic grammars

In these cases, some grammars are defined using rules that show how some composed categories may be decomposed in a succession of categories (that can be simple or composed). Each rule has a weight that refers its probability of apparition. There is also a table that stores the probability a syntactic category of a root has to start a sentence. The sentences are parsed using the *chart* bottom-up method [Allen-95].

A.- Morpheme prediction

In this system, the dictionaries continue storing syntactic information. The system offers the most probable roots with the appropriate starting whose categories are most probable to appear after the categories of the beginning of the sentence; then, the suffixes are proposed in a similar way and the process is repeated again and again. With the first word of a sentence, the most probable roots are proposed taking into account the probability table. As can be seen, the entire starting of the sentence is considered (so the computational complexity of this approach is higher than the previous ones).

In the adaptation, there is a need to update the frequencies of the dictionaries, the weights of the rules and the entries of the probability table.

B.- Morpheme-based prediction with acceptation of entire words

In our studies, we have found it is very usual in Basque that a root acts like an entire word (that is to say, no suffix is added to a root very frequently). This approach is based in this feature in order to enhance the results the previous approaches achieve.

The lexicons, the rules and the probability table are the same than the last previous approach. It tries to guess the entire word, if it is possible; if it is not, the roots are tried and then the suffixes as in the case of that approach. So, as can be seen, the entire beginning of the sentence is taken into account again.

The adaptation is also the same than in the last previous approach.

In order to distinguish if a root is accepted as a word or a mere root, the acceptation protocol is changed to make possible the total and partial acceptations, respectively.

3.- Achieved results

These five presented approaches have been evaluated by programs that emulate the user's behaviour [Garay-01]. In every approach, the results that are achieved when the system offers from 1 to 10 proposals to the user are studied. We have limited the study up to 10 proposals because of the difficulties to manage more than 10 proposals in the physical interface: it is complicated, very exhausting and time consuming to read them and select one of them.

The word prediction approach is associated with a 8,969 entries (composed each by a word and its frequency) lexicon that it has been created starting from some Basque corpora. The lexicon of the rest of the approaches is also taken from those corpora decomposing words in their roots and suffixes. In concretion, they have 4,771 roots and 233 suffixes (44.21% items less than in the first approach).

The study of the performance of the techniques have been done by using some texts that have not been used while creating the lexicons, to prevent favourable tendencies. In this paper we show the results achieved with the most representative text. Because there is a need to distinguish roots and suffixes in four of the approaches while this is not necessary in the first one, we have two versions of that text: one with delimiters of roots and suffixes (for the last four approaches) and the other without the delimiters (for the first one), as can be seen in figure 1. So, depending on the used version the text has 2,691 or 2,974 characters. The text is a technical report related about computer science and Internet.

Atal hau aipatutako zerbitzuak mantendu eta antolatzeaz arduratuko da, baita erabiltzaile interesatuei beren informazioa unibertsitateko intranet sarean eta interneten eskueran izateko aholkatzeaz ere.

Atal^ hau^ aipa^tutako zerbitzu^ak manten^du eta^ antola^tzeaz ardura^tuko da^ , baita^ erabiltzaile^ interesatu^ei beraiek^en informazio^a unibertsitate^ko intranet^ sare^an eta^ internet^en eskuera^n izan^teko aholkatze^az ere^.

Figure 1. Example of a Basque sentence without delimiters of morphemes (above) and with them (below), being "^" the delimiter.

In the figure 2, we show the keystroke savings[3] in percentiles that we have been obtained in the different approaches with that text. The main qualitative enhancements appear when the morphemes are separately treated (instead of treating the words) and when a root may be accepted as a whole word (instead of not giving this possibility). The second approach enhances the results of the first one between 5.83% and 11.08% in absolute values (28.24%-35.89% in relative values). The third presented approach enhances 1.55%-2.66% in absolute values (4.80%-6.50% in relative values) the results of the second one. The fourth one is better than the third one in 0.27%-1.11% absolute (0.62%-3.15% relative) and the last one enhances 2.63%-3.13% in absolute way the fourth one (6.87%-9.16% relative enhancement).

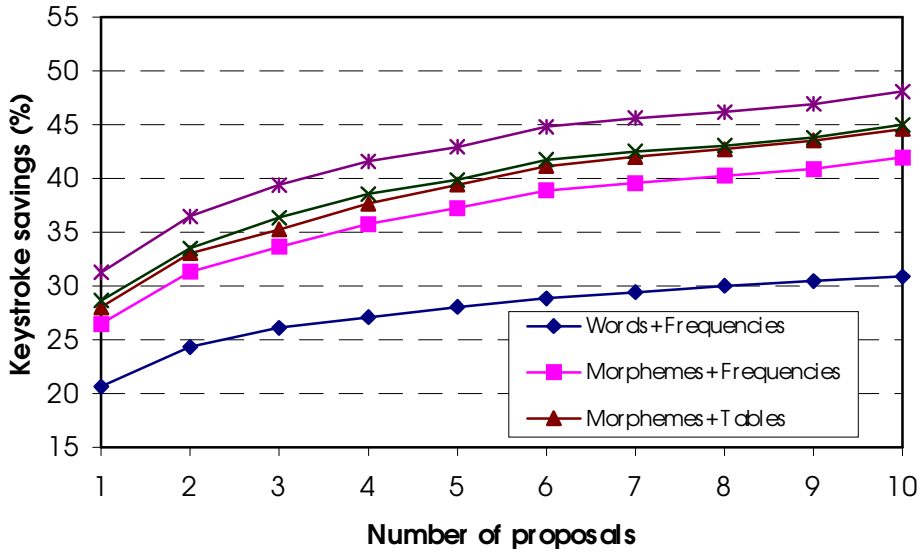

Figure 2. Keystroke savings with the different approaches.

In figure 3 the hit ratios[4] that have been achieved are shown. In general, the last presented approach achieves the best hit ratios, with an increase in absolute terms that vary between

[3] *Keystroke savings are defined as:* $\frac{total\ \#\ of\ keystrokes\ -\ \#\ of\ written\ keystrokes}{total\ \#\ of\ keystrokes}$

[4] *The hit ratio in word prediction is:* $\frac{\#\ of\ times\ a\ word\ is\ guessed}{\#\ of\ written\ words}$. *When distinguishing roots and suffixes hit ratio (HR) is calculated as* $HR = \frac{HRR + HRS}{2}$, *being HRR the hit ratio for roots and HRS the hit ratio for suffixes. In the last approach, if HRW is the hit ratio for roots that work as words, P the probability that there is no suffix added to a root*

12.75% and 15.23% in absolute values the next one (generally the fourth one). This enhancement with respect to the rest is mainly due to the fact that when accepting a root as a word there is an implicit acceptation of its null suffix, and this type of suffix is otherwise impossible to guess (at least in the approaches that distinguish roots and suffixes). The approaches that work with morphemes are similar each one with the rest. With one proposal, the word prediction presents a similar hit ratio, but in the rest of the cases is the worst one.

4.- Conclusions

In this paper we have presented five approaches to make prediction in an inflected language such as the Basque. The second, third and fourth approaches are very similar, but the savings are enhanced in the order they have been presented. The last presented method is the fourth one with the possibility of accepting roots as words, due to the special characteristics of the Basque (where this fact has a high probability). The last approach is the one that achieves the best results. So, the same strategy of accepting roots as entire words may be applied to the second and the third presented approaches in order to enhance their results.

We suggest the last presented approach (morpheme-based prediction using syntactic grammars with acceptation of entire words) as the most appropriate to get the best results. Anyway, we also have to conclude saying the word prediction by frequencies is language-independent while the rest depend on the characteristics of the analysed language (in our case, the Basque).

Acknowledgements

This paper is part of the Ph. D. thesis memory presented by the first author in the University of the Basque Country. The involved work has been financed by the Basque Government, the Department of Economy of the local administration called Gipuzkoako Foru Aldundia, the University of the Basque Country and the PITER Program, supported by the CICYT and IMSERSO.

Authors will thank to people that help them to do their work, specially the people from the Laboratory of Human-Computer Interaction for Special Needs, in the Computer Science Faculty of the University of the Basque Country.

(1•P•0) and HR, HRR and HRS are previously defined: $HR = HRW * P + \left(\frac{HRR + HRS}{2} \right) * (1 - P)$.

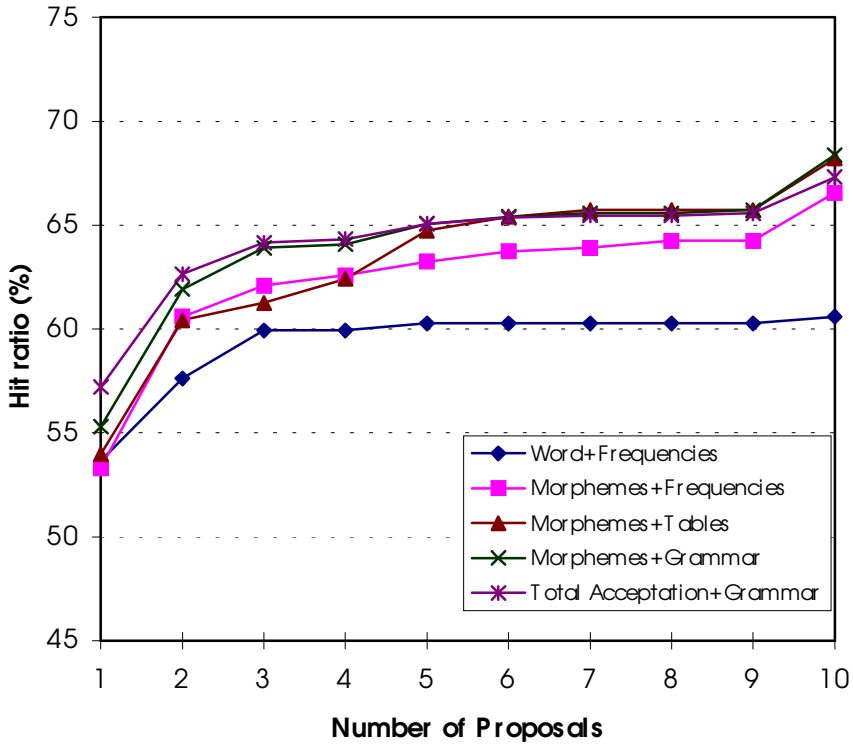

Figure 3. Achieved hit ratios in percentiles.

References

Agirre E, Alegria I, Arregi X, Artola X, Díaz de Ilarraza A, Maritxalar M, Sarasola K, and Urkia M. 1992. XUXEN: A Spelling Checker/Corrector for Basque Based on Two-Level Morphology. *Proceedings of the 3rd Conference on Applied Natural Language Processing*. Association for Computational Linguistics. 119-125.

Allen J. 1995. Natural Language Understanding. *Benjamin Cummings Publishing Company*. 2nd Edition.

Alm N, Arnott J L, and Newell A F. 1992. Prediction and Conversational Momentum in an Augmentative Communication System. *Communications of the ACM* 35, Number 5. 46-57.

Darragh J J, Witten I H and James M L. 1990. The Reactive Keyboard: A Predictive Typing Aid. *Computer*, 23 (Number 11 November). 41-49.

Garay N and González-Abascal J. 1994. Using Statistical and Syntactic Information in Word Prediction for Input Speed Enhancement. *Information Systems Design and Hypermedia*. Cépaduès-Éditions, Toulouse (France). 223-230.

Garay-Vitoria N and Abascal J G. 1997. Word Prediction for Inflected Languages. Application to Basque Language. *Proceedings of the Natural Language Processing for Communication Aids Workshop Sponsored by the Association for Computational Linguistics*. Association for Computational Linguistics. 29-36.

Garay-Vitoria, N. 2001. Sistemas de Predicción Lingüística. Aplicación a Idiomas con Alto y Bajo Grado de Flexión, en el Ámbito de la Comunicación Alternativa y Aumentativa. *Servicio Editorial de la UPV-EHU, Leioa (Bizkaia)*. ISBN: 84-8373-355-2.