Word Prediction for Inflected Languages. Application to Basque Language

Nestor Garay-Vitoria
Informatika Fakultatea
Euskal Herriko Unibertsitatea
649 Postakutxa
E-20080 Donostia (Basque Country)
E-mail: nestor@si.ehu.es
Phone: +34 43 21 80 00
Fax: +34 43 21 93 06

Julio G. Abascal
Informatika Fakultatea
Euskal Herriko Unibertsitatea
649 Postakutxa
E-20080 Donostia (Basque Country)
E-mail: julio@si.ehu.es
Phone: +34 43 21 80 00
Fax: +34 43 21 93 06

Abstract

Several word prediction methods to help the communication of people with disabilities can be found in the recent literature. Most of them have been developed for English or other non-inflected languages. While most of these methods can be modified to be used in other languages with similar structures, they may not be directly adapted to inflected languages. In this paper some word prediction techniques are reviewed and the difficulties to apply them to inflected languages are studied. Possibilities for word prediction methods that cope with the enormous number of different inflexions of each word are proposed, using Basque as the target language. Finally, conclusions about word prediction for inflected languages are extracted from the experience with the Basque language.

1 Introduction

So far, word prediction methods have been developed in order to increase message composition rate for people with severe motor and speech disabilities. These methods try to guess what is going to be the current or even the next word the user is trying to type. Their results are normally measured in terms of keystroke savings (Ks)[1].

To our knowledge, the design of word prediction methods is mainly focused on non-inflected languages, like English. Words in these types of languages have a small amount of variation, like the ones due to number (singular or plural) for instance,

---

[1]
$$Ks = \frac{Nchar - Nkeystrok}{Nchar}$$

house/houses, spy/spies. Some other languages admit differences in gender, for example in French: *voisin/voisine.* When the number of different forms of a word is small, it is possible to include all of them in the dictionary used in word prediction. Nonetheless, inflected languages can have a huge number of affixes that determine the syntactic function of each word and therefore it is not possible to include every variation of a word in the dictionary. So, other methods have to be tried for languages that use extensively prefixes, infixes or suffixes.

As a starting point, let us show what the declension of a word in Basque may be, by means of an example. The declension of the dictionary entry *mendi* (which means "mountain") can be seen in Table 1. This table is valid only for words referring to objects, but there are different tables for declensions of words referring to living beings. Whether the last letter of the lemma is a vowel or a consonant, different tables of declensions are also used. As shown, there are sixty-two possible word-forms for a single dictionary entry. In addition, most of the cases admit the recursive concatenation of suffixes. So, the number of possible cases grows. It has been estimated that nouns may mathematically have even 458,683 inflected forms in Basque language, taking into account two levels of recursion, (Agirre et al., 1992).

There also are other suffixes which are not shown in Table 1, as those applied to a verb for subordinate sentences.

Even if prefixes and infixes are possible the Basque language is declensed mainly by suffixes. There are some prefixes that can be used in some specific cases (for example, a prefix for verbs may indicate the absolutive case in the sentence), but in general their frequency of apparition is not very relevant. The same thing happens with the infixes: there are few of them in Basque and their frequency is not very relevant. Their prediction makes sense mainly if the

	Indefinite	Definite Singular	Definite Far Plural	Definite Near Plural
Absolutive	mendi	mendia	mendiak	mendiok
Ergative	mendik	mendiak	mendiek	mendiok
Dative	mendiri	mendiari	mendiei	mendioi
Possessive Genitive	mendiren	mendiaren	mendien	mendion
Comitative	mendirekin	mendiarekin	mendiekin	mendiokin
Benefactive	mendirentzat	mendiarentzat	mendientzat	mendiontzat
Motivative	mendirengatik	mendia(ren)gatik	mendiengatik	mendiongatik
Instrumental	mendiz	mendiaz	mendiez	mendioz
Inessive	menditan	mendian	mendietan	mendiotan
Ablative	menditatik	menditik	mendietatik/rik	mendiotatik/rik
Allative	menditara	mendira	mendietara	mendiotara
Allative of Destination	menditaraino	mendiraino	mendietaraino	mendiotaraino
Allative of Direction	menditarantz	mendirantz	mendietarantz	mendiotarantz
Local Genitive	menditako	mendiko	mendietako	mendiotako
Destinative	menditarako	mendirako	mendietarako	mendiotarako
Partitive	mendirik	———	———	———
Prolative	menditzat	———	———	———

Table 1: Declension of *mendi* (Kintana et al., 1988).

word is an auxiliary or a declined verb. For the rest of the cases, it seems better to treat the affix in combination with the lemma as a new lemma, if this combination is usual. Doing this, the complexity of operations decrease because there is only the need to treat lemmas and suffixes.

Thus, in this paper, the problem of suffixes will mainly be mentioned, because our target language is the Basque language.

2 Word Prediction Methods for Non-Inflected Languages

In this section some of the methods that have been used in word prediction for non-inflected languages are summarised. This small review will serve as a basis for coming sections in order to identify the key aspects that are involved in prediction. These methods are going to be presented by increasing complexity, from the simplest to the most complex.

2.1 Probabilistic Methods

2.1.1 Word Prediction Using Frequencies

The simplest word prediction method is to built a dictionary containing words and their relative frequencies of apparition. When the user starts typing a string of characters α the predictor offers the n

most frequent words beginning by this string in the same way they are stored in the system. Then, the user can choose the word in the list he or she wanted to enter or continue typing if it is not in the list. There are several studies about word frequencies in different languages, for instance (Beukelman et al., 1984) gives information about the frequency of word occurrence in English used by some disabled people.

If the dictionary does not contain inflected words (that is, if there are just the lemmas), it may need some correction by the user (or by the system) in order to adjust its concordance with other related words. For instance, it may need to adjust the gender: "*C'est une voiture* fantastique" or the number "*A lot of* cars". The dictionary uses to be an alphabetically ordered list of words and their frequencies, but other possible dictionary structures can be found in (Swiffin et al., 1987a). This prediction system can be adapted to the user by updating the frequency of the word in the dictionary each time this word is used. Words seldom employed can be replaced by others which are not in the dictionary. The inclusion of new words is not difficult because all the information that is required is their frequency. Further information about this type of prediction can be seen in (Colby et al., 1982), (Garay et al., 1994a), (Heckathorne et al., 1983), (Hunnicutt, 1987), (Swif-

fin et al., 1987a), (Venkatagiri, 1993).

To enhance the results of this method, an indication about the "recency" of use of each word may be added. In this way, the prediction system is able to offer the most recently used words among the most probable ones beginning by α. Each entry in the dictionary is composed by a word, its frequency and its recency of use. Adaptation of the dictionary to the user's vocabulary is possible by updating the frequency and recency of the each word used. (Swiffin et al., 1987a) observes that this method produces small better savings in the number of keystrokes needed than in the previous approach, but more information must be stored in the dictionary and the complexity is also increased.

2.1.2 Word Prediction Using Probability Tables

Another possibility is to use the relative probability of appearance of a word depending on the previous one. To implement this system a two-entries table is needed to store the conditional probability of apparition of each word W_j after each W_i. If the dictionary contains N words the dimension of the table will be of $N*N$. That is, it will have N^2 entries, but most of the values in the table will be zero or close to zero.

In some cases it could be possible for the system to give proposals before entering the beginning of a word. The recency of use may also be included in this approach. This method is hardly adaptable to include the user preferred words because the dimensions of the table cannot be changed. This difficulty leads to the design of modified versions, like the one that uses only the most probable pair of words, reported as in (Hunnicutt, 1987).

2.2 Syntactic Word Prediction

2.2.1 Syntactic Word Prediction Using Probability Tables

This approach takes into account the syntactic information inherent to the languages. To this end two kinds of statistical data are used: the frequency of apparition of each word and the conditioned probability of each syntactic category to follow every other syntactic category. In this way, the set of words that are candidates to be proposed by the predictor is restricted to the ones that match the most probable syntactic role in the current position of the sentence, thus increasing the hint rates. This syntactic table is smaller than the one used in the previous approach, and the proportion of probabilities which are close to zero is also smaller. Each entry in the dictionary will associate a word with its syntactic category, and

its frequency of apparition. Words can be sorted by syntactic categories to facilitate the selection process. When a word is syntactically ambiguous, that is, when more than one category is possible for a given word, one entry for each possible category may be created. The table of conditional probabilities of syntactic categories has a fixed size and it is built before the use of the predictor. Adaptation to the user's lexicon is possible because there is no need to increase the size of the table. New words are included in the dictionary with a provisional syntactic category deducted from its use. Later on, the system may require some help from the user to verify if the categorisation was correct. It could be also possible to add some morphological information in the dictionary to propose the words with the most appropriate morphological characteristics (gender, number). This could increase the hint rate of the predictor. Some systems that use this approach are described in (Garay et al., 1994a) and (Swiffin et al., 1987b).

2.2.2 Syntactic Word Prediction by Using Grammars

In these approaches, the current sentence is being parsed using a grammar to get the most probable categories. Parsing methods for word prediction can be either "top-down" (Van Dyke, 1991) or "bottom-up" (Garay et al., 1994b), (Garay et al., 1997). So, there is a need to define the syntactic rules (typically $LEFT <- [RIGHT]+$, usually being $LEFT$ and $RIGHT$ some syntactic categories defined in the system) that are used in a language. Within a rule, it could be possible to define concordance amongst the components of the right part (either in gender and/or in number). Then, the proposals may be offered with the most appropriate morphological characteristics. It is necessary to leave open to the user the possibility of changing the word's ending. For example, if there is a mismatch in the rule used by the system, it may be necessary to modify the end of an accepted proposal. The dictionary is similar to the one used in the previous approach with the addition of morphological information to allow concordance. The complexity of this system is also larger because in this case, all the words of the sentence that appear before the current word are taken into account, while in the previous approaches only one previous word was used. The adaptation of the system for the new words is made increasing the word frequencies and the weights of the rules. The inclusion of new words is similar to the one in the previous approach. The use of grammars for word prediction is also shown in (Hunnicutt, 1989), (Le Pévédic et al., 1996), (Morris et al., 1991) and (Wood et al., 1993).

31

2.3 Semantic Word Prediction

These methods are not very used, because their results are similar to those of the syntactic approaches, but the increase in complexity is great. Maybe the simplest method that can be used is the semantic word prediction by using parsing methods. In this approach each word has some associated semantic categories, while in the previous one categories were purely syntactic. The rest of the features (the procedure, complexity, structure of the dictionary, adaptability...) are similar to the previous one. Nevertheless, the problem of giving semantic categories to the words is very complex and it results difficult to be programmed. Some authors propose semantic categorisation made "by hand" (Hunnicutt, 1989).

There may be other methods to treat the semantic information, but their complexity is going to be very great for a real-time system as the word predictors are intended to be, even the time requirements (maybe a few seconds between two consecutive keystrokes of an impaired person) are not very strong for the computational capacities of today's equipment.

3 Application of Mentioned Word Prediction Methods to Inflected Languages

In this section the use of previously reviewed word prediction methods for non-inflected languages is studied and their suitability for inflected languages is discussed. So, the key question is: Are the word prediction methods that we have previously shown useful for inflected languages?

As we mentioned in the introduction, in non-inflected languages it is feasible to include in the dictionary all the forms derived from each lemma, taking into account that the number of variations is quite small. For instance, in English *friends* is the only variation (without creating composed words) of *friend*, and the verbs have a few variations too. In Spanish, the word *amigo* (with the same meaning than *friend*) may vary in gender and number, giving the words: *amiga, amigos* and *amigas*. But the variations that the word *adiskide* (same meaning as *friend* or *amigo*) may have in Basque make it impossible to store them in the dictionary of the system. This is one of the changes to be taken into account for the design of a predictor for this type of languages. In inflected languages, the complexity in making the changes is very high, because of the number of possibilities. One possibility is to group the suffixes depending on their syntactic function to make it possible to have an easy automatisation. In

addition, we shouldn't forget that suffixes may be recursively concatenated.

In the previously presented prediction methods, the ones using probabilistic information mainly work with the words as isolated entities. That is, they work seeing each word in the dictionary as a whole to be guessed, without taking into account the morphosyntactical information inherent to the languages. So, a word that is not at the lexicon can not be guessed. The impossibility to store all the combinations of a word, make these methods not very suitable for inflected languages[2].

Therefore, it would be very interesting to treat the entire sentence. Then, the first syntactic approach is not very useful, because it only takes into account the previous word. And the second one is very hard to implement, because of the number of variations a word may have. Maybe a great number of rules have to be defined to cope with all the variations, but in this way the probabilities to guess the rule which is being used are very small, because of their variety.

The same thing happens with the semantic approach, which has, as it has been said before, the same procedural characteristics as the second syntactic one.

So, the complexity needed to create a correct word, including all the suffixes it needs, in inflected languages may make it necessary to search for other prediction methods, apart from all that were shown in the previous section.

---

[2]To know what the suitability for the next shown approaches can be, let us show a special case for Basque: verbs, mainly auxiliary verbs. They depend not only on the subject (which normally appears as absolutive or ergative cases) but also on the direct complement (if the sentence is transitive this complement has the absolutive case while the subject has the ergative case) and on the indirect complement (the dative case). For instance, the auxiliary *dizut* is related to the subject of the first person singular, the object complement in the singular and the indirect complement of the second person singular. But if the subject is in the third person plural, the indirect complement in the first person plural and the direct complement is in the plural, the needed auxiliary has to be *dizkigute*. Both cases are in the present of the indicative. If the tense of the verb changes, the verb itself also changes (for example, the past of the indicative of *dizut* is *nizun* and the past of *dizkigute* is *zizkiguten*). There also are some cases in which the verb depends on the gender of the absolutive, ergative or dative cases.

4 Word Prediction in an Inflected Language. Application to Basque Language

4.1 First Approach to Solving the Prediction Problem in an Inflected Language

As we have seen in the previous section, it is very difficult to predict complete words in inflected languages because of the variations a word may have. As there is a huge variety of inflected languages, let us concentrate on the particular characteristics of the Basque language, customising to this case the operational way.

For this first approach, due to the above mentioned primacy of suffixes (over other affixes) in the Basque language, and to simplify the problem, prediction in Basque is divided in two parts: prediction of lemmas and prediction of suffixes. Thus, two dictionaries (one for lemmas and other for suffixes) are used. The first one includes the lemmas of the language alphabetically ordered with their frequencies and some morphologic information in order to know which possible declensions are possible for a word. The second one includes suffixes and their frequencies ordered by frequencies.

To start the prediction, the system tries to anticipate the lemma of the next word. Most of the methods seen in previous sections can be used for this purpose. When the lemma is accepted (or typed entirely if the predictor fails), the system offers the suffixes that are correct for this lemma ordered by frequencies. As the acceptable suffixes for a noun can be about 62 (as we have seen in the Table 1) only the most probable n suffixes are offered[3]. As can be seen, the operational way is very similar to word prediction using tables of probabilities, but there is some added complexity because the system (and also the user) has to distinguish between lemmas and suffixes. In addition, more than one table of probabilities may be necessary to properly make predictions. Apart from the increase of the complexity, a decrease of the keystroke savings may be expected, because of the need of accepting at least two proposals for completing a word (while at least only one proposal is required with predictors for non-inflected languages).

Even if some promising results have been obtained, there are still some problems to solve in this approach.

- First of all, due to the possibility of recursively composed suffixes (concatenating the existing

---

[3]With n depending on the interaction method

ones) the system has to again propose a list of suffixes until the user explicitly marks the end of the current word (maybe inserting a space character).

- The recursive behaviour is one of the reasons to create more than one table of probabilities which stores the probability of apparition of a suffix immediately after the previous one.

- The system may be adapted to the user updating the frequencies in the lexicons and the probabilities of the tables. To include a new lemma in the dictionary, it is necessary to obtain its morphological characteristics.

- Finally, due to the special characteristics of the verbs (that include any kind of affixes in concordance with other words in the entire sentence) their prediction requires a special treatment.

Therefore, it seems interesting to do a syntax approach for these types of languages, because otherwise, the problems of this approach are very difficult to solve.

4.2 Second Approach to Solving the Prediction Problem in an Inflected Language

This approach will try to alleviate the above mentioned problems. The lemmas and the suffixes are still treated separately, but syntactical information is included in the system. This can be done by adding syntactic information to the entries of the dictionary of lemmas, and some weighted grammatical rules on the system. The main idea is to parse the sentence while it is being composed and to propose the most appropriate lemmas and suffixes. In principle, the parsing allows storing/extracting the information that has influenced in forming the verb. There exist systems that verify the morphologic and syntactical correctness of a Basque sentence, but the complexity of the Basque verb avoids its anticipation. To face this problem, the most frequent verb forms are included in the dictionary, and a morphological generator permits their modification or the addition of suffixes when it is necessary.

As there are no probability tables, there is no problem related to their extension. The adaptation of the system is made by updating the frequencies of the lemmas and suffixes and the weights of the defined rules. The inclusion of a new lemma in the lexicon might cause some lack of syntactic information. To solve this problem, there are some possibilities. First, the predictor tries to guess the category, depending on the most highly weighted rule

at that point of the sentence. Second, the predictor asks the user directly about the information. The first approach can produce false assumptions, while the second one slows the message composition rate and demands a great knowledge of the syntax by the user. There is another possibility: the predictor marks the lemma and the user is asked to complete the needed information after ending the session.

Finally, recursion may be included into the defined rules. Most of the grammars may have an implicit recursion which may be shown by rules. For instance, let us consider these rules:

NP <- $Noun$ PP

PP <- $Prep$ NP,

where NP means Noun Phrase, PP, Prepositional Phrase, $Noun$ is a noun and $Prep$, a preposition. As can be seen, these rules can be expanded to:

NP <- $Noun$ $Prep$ NP.

So, the NP is on the left and on the right of the same rule, and a recursion happens. This recursion may be used as a way to indicate the recursion of the concatenation of the suffixes, because they can express the syntactic role of a word in a sentence, as it was noted in the introduction.

The operational way and the order of complexity are similar to the word prediction using grammars. Nevertheless higher complexity may be expected mainly due to the existence of lemmas and suffixes. So, poorer keystroke savings are expected.

To enhance this approach, it seems interesting to try to guess the entire word, that is, a lemma and its associated suffix. This system will be easier to use (there is no need to force users to know what the lemma and what the suffix of a word are) and may have better results, measured in terms of keystroke savings.

4.3 Third Approach to Solving the Prediction Problem in an Inflected Language

Taking into account the previous experience, a third approach could be tried. Built as a combination of the previous ones, the main idea is to guess the entire current word. It treats the beginning of the sentence like the first approach, using statistical information. While advancing in the composition of the sentence, the system parses it and uses this information to offer the most probable words, including both lemma and suffix, like the second approach does. The first word of the sentence is treated using the first approach seen. But to minimise the problems related to that approach, the rest of the sentence is treated using the second approach.

In this way, only three tables would be needed:

one with the probabilities of the syntactic categories of the lemmas to appear at the starting of a sentence, another with the probabilities of the basic suffixes to appear after those words and the third with the probabilities of the basic suffixes to appear after another basic suffix (and to make possible the recursion). All of these tables would have fixed sizes even when new lemmas are added to the system. The adaptation of the system would be made updating the first table and, while the suffixes would be added to the word, the other two tables would be also updated. With relation to the new lemmas that do not have the information completed, they might update, or not, the first of the tables if a entry for the unknown cases is included; otherwise they would remain unchanged. Finally, the problem of verb formation in Basque is not solved and the most frequent verb forms are included in the dictionary in the same way as in the second approach.

5 Conclusions

Our experience with the Basque language in word prediction applied to Alternative and Augmentative Communication for people with disabilities, shows that prediction methods successful for non-inflected languages are hardly applicable to inflected ones. The high number of inflexions for each word makes their inclusion in the lexicon impossible. Different approaches have been studied to overcome this problem. To be able to predict whole words it is necessary to determine the syntactic role of the next word in the sentence. That can be done by means of a syntactic analysis "on the fly". Nevertheless the results of the evaluation of these methods with the Basque language are not as good as the ones obtained with non-inflected languages.

6 Acknowledgements

References

Eneko Agirre, Iñaki Alegria, Xabier Arregi, Xabier Artola, Arantza Diaz de Ilarraza, Montse Maritxalar, Kepa Sarasola, and Miriam Urkia. 1992. XUXEN: A Spelling Checker/Corrector for

Basque Based on Two-Level Morphology. In *Proceedings of the 3rd Conference on Applied Natural Language Processing*, pages 119–125, Trento, Italy, March-April. Association for Computational Linguistics.

David R. Beukelman, Kathryn M. Yorkston, Miguel Poblete and Carlos Naranjo. 1984. Frequency of Word Occurrence in Communication Samples Produced by Adult Communication Aid. *Journal of Speech and hearing Disorders*, Vol. 49: 360–367.

Kenneth M. Colby, Daniel U. Christinaz, Roger C. Parkison, Mark Tiedemann. 1982. Predicting Word-Expressions to Increase Output Rates of Speech Prosthesis Used in Communication Disorders. In *IEEE Proc. ICASSP-82*, pages 751–754.

Nestor Garay and Julio González Abascal. 1994. Using Statistical and Syntactic Information in Word Prediction for Input Speed Enhancement. In *Information Systems Design and Hypermedia* edited by Claude Chrisment, pages 223–230. Cépaduès-Éditions. Toulouse (France).

Nestor Garay-Vitoria and Julio González Abascal. 1994. Application of Artificial Intelligence Methods in a Word-Prediction Aid. In *Computers for Handicapped Persons* edited by W. L. Zagler, G. Busby and R.R. Wagner, pages 363–370. Springer-Verlag. Berlin.

Nestor Garay-Vitoria and Julio González Abascal. 1997. Intelligent Word-Prediction to Enhance Text Input Rate. In *Proceedings of the Intelligent User Interfaces 97 International Conference*, pages 241–244, Orlando (Florida), January. ACM Press, New York.

Craig W. Heckathorne, Dudley S. Childress. 1983. Applying Anticipatory Text Selection in a Writing Aid for People with Severe Motor Impairment. *IEEE MICRO*, June, pages 17–23.

Sheri Hunnicutt. 1987. Input and Output Alternatives in Word Prediction. *STL/QPSR 2-3/1987*, pages 17–29. Dept. of Speech Comm. and Music Acoustics. Royal Institute of Tech. Stockholm.

Sheri Hunnicutt. 1989. Using Syntactic and Semantic Information in a Word Prediction Aid. In *Proceedings of the European Conference on Speech Communication and Technology "Eurospeech 89"*, Vol. 1: 191–193. Paris, September.

Xabier Kintana, Jon Aurre, Ricardo Badiola, Sabin Egileor, Juan Luis Goikoetxea, Juanjo Zearreta, Koldo Zuazo, **Ekiten** taldeko irakasleak. 1988. Hiztegia Bi Mila: Euskara-Espainiera/Espainiera-Euskara. *ELKAR*, Donostia, Basque Country.

Brigitte Le Pévédic and Denis Maurel. 1996. Un Logiciel d'Aide a la Communication pour des Personnes Handicapeés. In *Proceedings of the NLP+IA '96 (TAL+AI '96)*, pages 211–216. Moncton, N. B., Canada.

Corinne Morris, Alan Newell, Lynda Booth and John Arnott. 1991. SYNTAX PAL-A System to Improve the Syntax of Those with Language Dysfunction. In *Proceedings of RESNA 14th Annual Conference*, pages 105–106, Kansas City.

Andrew L. Swiffin, John L. Arnott, J. Adrian Pickering and Alan F. Newell. 1987. Adaptive and Predictive Techniques in a Communication Prosthesis. *Augmentative and Alternative Communication*, pages 181–191.

Andrew L. Swiffin, John L. Arnott and Alan F. Newell. 1987. The Use of Syntax in a Predictive Communication Aid for the Physically Handicapped. In *Proceedings of RESNA 10th Annual Conference*, pages 124–126, San Jose, Ca.

Julie A. Van Dyke. 1991. Word Prediction for Disabled Users: Applying Natural Language Processing to Enhance Communication. Thesis submitted for the degree of Honors Bachelor of Arts in Cognitive Studies. University of Delaware, June.

Horabail S. Venkatagiri. 1993. Efficiency of Lexical Prediction as a Communication Acceleration Technique. *AAC*, Vol. 9: 161–167.

Matthew E. J. Wood, Eric Lewis. 1993. Grammatical Recognition in Computer Aided Conversation. In *Proceedings of the European Speech Communication Association (ESCA) Workshop on Speech and Language Technology for Disabled People*, pages 115–118. Stockholm, Sweden, May-June.