

VITIPI : A universal writing interface for all.

Philippe BOISSIERE Daniel DOURS

IRIT – UPS
118 Route de Narbonne
31 062 TOULOUSE Cedex (France)
TEL : (33) (0)5 61 55 63 41 FAX : (33) (0)5 61 55 62 58
e-mails : {[boissier](mailto:boissier@irit.fr), [dours](mailto:dours@irit.fr), [vigouroux](mailto:vigouroux@irit.fr)}@irit.fr
[http :www.irit.fr](http://www.irit.fr)

INTRODUCTION :

VITIPI's aim is to increase speed text acquisition in all computer applications. Unlike the others systems, it doesn't display lists of unexpected words but provides the ending of words without end-user's intervention. When VITIPI faced with an unknown or unexpected word, containing typing errors or orthographic mistakes, it can continue to predict the end of the word. It takes into account previous words for prediction. VITIPI database is made up with user's corpus so that it is well adapted to user's vocabulary and grammar. It will not be linked to a devoted word processor, it will could run on all windows applications.

If we only consider isolated French words, VITIPI provides *26 % of predicted letters* on a vocabulary size 5,930 words. *72 % of typing errors* are corrected and *75 % of orthographic mistakes*.

Hence, if these results seemed good they could be better. It could be notice that there is a relation between vocabulary size and ratio prediction (*predicted letters divided by total letters*). The more grows the vocabulary size, the more it will be faced with ambiguities, and the ratio prediction goes down. It will have to wait user's letters to clear up ambiguities. Inversely, the more decreases the vocabulary size, the more goes up the ratio prediction. If we want to increase ratio prediction, we have to reduce the vocabulary size without hindering the user potential. It is important to outline that after words succession, a very small set of words can appear. Our system uses this property to build the set of words that could be wrote at this very instant.

VITIPI SYSTEM

VITIPI system creation is based on two cooperated process. The first one intends to create the set of the words that could be written at this very instant. We have chose *N-gram* to implement it. The second one tries to predict letters of the words that have been selected by the first process. Transducers have been chose to implement it.

This approach has been modelised with N-gram pattern. Thanks to it when a word had been written, VITIPI can find in the corpus the set of following words. If we take into account the last two wrote words, the set of following words is reduced. If now we consider the last n-1 wrote words, the number of following words is highly reduced and could tend towards one, or maybe no word. This set of following words that appear after a n-1 words string can be selected to constitute set of words that are likely to write at this very instant. We are now going to examine how letters prediction is made in selected words.

In order that VITIPI automatically provides letters according to user inputs, lexical base is modelised by a transducer. A transducer transforms an input string into an output string. Input corresponds to user's typed letter, output is system written letters.

We have got to notice that in the system, we can find several transducers (*under-transducers*), containing various vocabularies. In an under-transducer, we can find the set of words following: one word, a pair of words, ... a string of n-1 words. To access to an *under-transducer*, we have to consider words of the text as an alphabet letters. Each word is associated to a single "word-number" assigned by order of appearance in the corpus. These "word-numbers" are considered as transducer inputs. Then, last words knowledge leads us to right under-transducer.

SYSTEM IMPLEMENTATION

We have developed a specialised transducer minimisation process that doesn't introduce new words or sentences in the system. When system is used, the words and/or sentences can belong or not to learning corpus. If it belongs, system skims through transducer with user word. If it doesn't belong to corpus learning, when system skims through transducer, it will be face with undefined transition. An inferred transition will then be done. Five procedures were developed to release system when it is faced to altered or new words. When user writes, he will necessary have to write sentences that don't belong to learning corpus. In this case two events can occur:

- It is a system unknown word and we come back to the following case. If it is really a new word, then it will necessary be a new sentence which at end, will be integrated to corpus.
- It is a system known word, as all anterior words, but words never appeared in this order in learning corpus. We have developed for such event, two new inference procedures.

VITIPI RESULTS

We made a first test with the weather forecast on the French web site of METEOFRACTANCE. We noticed that for the first forecast day, the system displays 14 % of output letters with isolated words. By the opposite, if it takes into account 10 previous words, it displays 16.5 % of output letters. For the 20th forecast day, system displays 23 % of output letters with isolated words. By the opposite, when 10 previous words are taken into account, system provides 41 % of output letters. If we don't use basic French vocabulary, system provides 44 % of output letters. All forecast sentences were unknown for the system. When all sentences belong to the corpus, the prediction rate goes up to 77%.

CONCLUSION

At the beginning VITIPI had been made for disabled people to increase speedwriting. But nowadays, in the information society, computers take a huge place, and if you are more than 40 or 50 years old, you may be out if you don't know how to use computer keyboards. Such people who have never learnt to use keyboards may become "keyboard disabled" and VITIPI interface will be very helpful for them.