FASTY – A Multi-lingual Approach to Text Prediction

Feature

Marco Baroni, *Austrian Research Institute for Artificial Intelligence (OeFAI), Vienna*

Written communication and information exchange is a vital factor in human society. Impairments, which lead to a reduction of typing speed, therefore, severely influence quality of life and cut off a person from equal participation in the information society.

This problem is being addressed by FASTY – a European project funded by the Fifth IST R&D Framework Programme. The project started in January 2001, and will run until December 2003. The consortium is led by Fortec – Vienna University of Technology (Austria). The other partners are the Austrian Research Institute for Artificial Intelligence (Austria), FTB (Forschungsinstitut Technologie-Behindertenhilfe, Germany), the Department of Linguistics at Uppsala University (Sweden), Multitel ASBL (Belgium), IGEL GmbH (Germany), Elisabethinum Axams (Austria), IkuT – Ingenieurbüro für Kunst und Technik (Germany), and Facultés Universitaires Notre-Dame de la Paix (Belgium).

Since languages display a high degree of redundancy, low-speed typists can be supported by Predictive Typing (PT) systems. Such systems attempt to predict subsequent portions of text by analysing the text already entered by the writer. Character-by-character text entry is replaced by making a single selection as soon as the desired word or sequence is offered by the system in the selection menu.

State-of-the-art programs for PT claim Keystroke Saving Rates (KSR) of up to 75%. This does not mean, however, that the text generation rate increases by a factor of four. Using PT consumes time, because the user needs to read the selection menu and make a decision. Only substantial KSRs will lead to an increase of communication speed. To double the text generation rate of a typical mouth-stick user, the program must offer a KSR of about 65%.

Such high rates are currently only achieved for English, a language almost unique in having a very limited set of inflectional endings. This property makes it ideally suited to the currently most popular PT technology, which uses a statistical approach based on the probability of word n-grams. By adapting programs designed for English to other languages (especially highly inflected ones), the KSR drops significantly (usually below 30%). Therefore, most motor/speech impaired persons will experience no gain in text generation rate from existing programs.

FASTY aims at providing impaired speakers of languages other than English with PT systems that perform as well as those that nowadays are available for this language only. FASTY is currently being implemented for Dutch, French, German, and Swedish, but it is based on a generic modular architecture, with a clear separation between processes and language-specific resources. This should make adaptation to other languages relatively easy.

The target languages of the FASTY project are highly inflecting. Depending on the syntactic context, words take different forms. As already mentioned, this makes standard n-gram language modelling techniques less effective. Thus, additional methods that are able to cope with syntactic constraints are needed. Furthermore, in most of FASTY's target languages (i.e., Dutch, German, Swedish), productively formed compounds are written as single orthographic strings (in contrast to English, where compound terms are groups of words separated by a blank character or, at least, a hyphen). This causes serious problems in terms of lexical coverage and data sparseness to systems that do not perform some type of compound processing.

The FASTY language component includes the following modules: word- and part-of-speech-based n-gram prediction; grammar-based prediction; compound prediction; morphological lexicon; user lexicon; collocation-based prediction. The modules are driven by a controller engine that manages the input requirements of each component, establishes the required input data from the context, and combines the outputs in a meaningful way.

Preliminary experiments indicate that the n-gram-based models, despite the problems mentioned above, still provide reasonable predictive power, and they constitute the core of the FASTY language component.

The grammar-based module performs a partial parse of the current input, and it ranks the predictions provided by the other modules on the basis of the grammatical information provided by the parse. Moreover, in contexts where all predictions by the core component are syntactically ill-formed, the grammar-based module generates well-formed predictions using the morphological lexicon.

The compound prediction module allows the user to type (nominal) compounds in multiple steps. The user can choose to complete the word constituting the first part of the compound (if it is in the prediction list), and then re-enter the prediction loop for the current word, now getting predictions for the second part of the compound (and this process can be repeated as many times as necessary to obtain longer compounds). >

Compound prediction is based on a set of compound-specific statistical models trained on a corpus where compounds have been split into modifiers and heads. In preliminary research to be reported at ECAI02 and COLING02, we have found that our split compound model leads to an improvement in compound word KSR of more than 15%, over a baseline model without compound analysis. However, integrating compound- and whole-word prediction so that compound completions do not get in the way of simple-word predictions turns out to be a rather difficult task, and we are still experimenting with alternative integration strategies.

The morphological lexicon provides the necessary morphosyntactic information to the grammar-based module. It also functions as a last resort prediction source, if all the other modules run out of completions before the user finds the word s/he intends to type.

The user lexicon is an additional, dynamic resource intended to support the style and vocabulary preferred by a particular user. It contains words and n-grams collected from the texts written by the user so far, and thus may contain words and phrases that may not be present in the general dictionary but are of importance for the user (e.g., names of people the user often addresses, specific terminology s/he is using, etc.) The user dictionary is automatically augmented during text entry, allowing the prediction of words new to the system after the first time they are used. At the end of the session the user can choose whether to save or discard new terms.

The final version of FASTY will also include a collocation-based module to provide predictions based on the degree of textual association between words (or word-classes). This module has not yet been implemented.

The language resources for all the FASTY languages are nearly ready, and we expect to complete the implementation of a realistic prototype of the whole system (excluding collocation-based prediction) very soon.

At this stage, the biggest issues are those pertaining to the integration of the various components of the system.

FOR INFORMATION

Marco Baroni is a researcher at the OeFAI, currently working at the FASTY project, and focussing in particular on compound prediction

Email: marco@ai.univie.ac.at

More about FASTY: www.fortec.tuwien.ac.at/fasty