in Miesenberger K., Klaus J., Zagler W. (eds.): Computers Helping People with Special Needs. 8th International Conference, ICCHP 2002, Linz, Austria, July 15-20, Proceedings, pp. 243-250. Lecture Notes in Computer Science, Vol. 2398, Springer, Berlin-Heidelberg-New York.
http://www.springer.de/comp/lncs/index.html

FASTY - A multi-lingual approach to text prediction

Johannes Matiasek[1], Marco Baroni[1], and Harald Trost[2]

[1] Austrian Research Institute for Artificial Intelligence,
Schottengasse 3, A-1010 Wien, Austria
{john,marco}@oefai.at
[2] Department of Medical Cybernetics and Artificial Intelligence,
University of Vienna
harald@ai.univie.ac.at

Abstract. Communication and information exchange is a vital factor in human society. Communication disorders severely influence the quality of life. Whereas experienced typists will produce some 300 keystrokes per minute, persons with motor impairments achieve only much lower rates. Predictive typing systems for English speaking areas have proven useful and efficient, but for all other European languages there exist no predictive typing programs powerful enough to substantially improve the communication rate and the IT access for disabled persons. FASTY aims at offering a communication support system significantly increasing typing speed, adaptable to users with different language and strongly varying needs. In this way the large group of non-English-speaking disabled citizens will be supported in living a more independent and self determined life.

1 Introduction

Whereas experienced typists will produce some 300 keystrokes per minute, persons with motor impairments achieve only much lower rates. One obvious alternative to manual typing is automatic speech recognition. However, diseases affecting the dexterity often also influence the ability to speak and/or the quality of vocal expression. Such persons have to rely completely on typing even for situations usually reserved for oral communication.

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 list.

1.1 Drawbacks in Existing Predictive Typing Systems

State-of-the-art programs for PT claim Keystroke Saving Rates (KSR) up to 75%. This does, however, not mean, that the Text Generation Rate (TGR) increases by the factor four. Using PT consumes time for reading the selection list

and making a decision. Only substantial KSRs will lead to an increase of communication speed. For every user there exists a (break-even) Keystroke Saving Rate below which the use of a Predictive Typing program will not result in an increase on the Text Generation Rate. For example, to double the TGR of a typical mouth-stick user the program must offer a KSR of about 65%.

For the English language even very simple PT programs yield acceptable Keystroke Saving Rates well above 30%. This success is due to the fact that - almost uniquely among European languages - English is a language with a very limited set of inflectional affixes. For this reason predictive typing became popular in English speaking countries quite early. This popularity kindled the development of more sophisticated PT programs in those countries. Presently, all available PT programs originate from the USA, Canada and the UK. A product search in spring 2000 identified 21 products from the USA, 6 from Canada and 5 from the UK but only one from Germany. When programs of English language origin are transferred to other languages (especially highly inflecting ones) the KSR drops significantly (usually below 30%). Therefore, most motor/speech impaired persons will experience no gain in TGR from existing programs.[3] To offer high KSRs a prediction program must be designed taking into account the properties of the language for which it is used.

1.2 The FASTY Approach

The FASTY system, by augmenting and improving the standard technology, aims at providing impaired speakers of languages other than English with PT systems delivering KSRs currently available for English only. FASTY is being developed for the German, French, Dutch and Swedish language, the concept, however, will be usable for most European languages. FASTY will assist motor, speech, learning and language impaired persons to produce texts faster, with less physical/cognitive load and with better spelling and grammar. FASTY will be configurable for different types of disabilities, different communication settings and different European languages. It will allow easier access to PC based office systems, to modern forms of IT communication and a faster usage of text to speech synthesizers for voice communication. FASTY is an intelligent system by using methods of Natural Language Processing, Artificial Intelligence, self adaptive user interfaces and knowledge bases in order to significantly increase KSR especially for those European languages which are highly inflected.

FASTY follows a generic approach in order to be multilingual by design. The language independent prediction software clearly separates between the predictor, and the language-specific resources. This will result in a system with potential application to many European languages without sacrificing performance.

[3] However, increasing speed is not the only reason for using PT. Some users prefer predictive typing because it is less stressful than character-by-character entry. Dyslexic and learning disabled persons benefit from the selection list, because it supports their efforts in word finding and improves their spelling.

Throughout development a user panel ensures the strong involvement of users (including primary end-users as well as pedagogues, therapists, carers and family members as secondary users) in the project.

The user interface design and the features of the predictor program aim at a wide coverage of end users (various disabilities) and secondary users (various roles in supporting the disabled person). Self adapting parameters and flexible configuring ensure a high degree of usability, user friendliness and accessibility. Innovative and ergonomic user interfaces for various existing input methods (standard keyboard, on-screen-keyboard, scanning) will be developed together with the predictor thus minimising time and effort for selecting the desired word from a selection list presented on the screen.

In this paper we will not go further into details of the user interface design and development but rather focus on the underlying language component.

2 Overall Architecture of the Language Component

The target languages of the FASTY project are highly inflecting ones posing additional challenges to word prediction. Depending on the syntactic context, words take different forms. This makes standard language modelling techniques employing n-gram models of word forms less effective. Thus, additional methods that are able to cope with syntactic constraints are needed. Furthermore, for most of FASTY's target languages (i.e., Dutch, German, Swedish) compounds are written as single orthographic strings, in contrast to English and French, where compound terms are groups of words, still separated by a blank character or at least a hyphen. Since compounding is very productive in all languages, this renders all attempts to have a complete lexical coverage of these languages hopeless.

All the prediction modules are driven by a controller engine that takes care of the input requirements of each module, establishes the required input data from the context, and combines the module's results in a meaningful way, yielding the desired predictions. The operation of the controller is adjustable by language- and user-specific parameters. While the core engine of FASTY is fully implemented the additional components and the controller are currently under development.

2.1 Core Components

N-gram-based Statistical Prediction Despite the problems with highly inflecting languages discussed above, preliminary experiments with German as well as experiences with a Swedish system [4] have shown that n-gram based methods still offer quite reasonable predictive power. Furthermore, the data sources needed by an n-gram based predictor, i.e., frequency tables of word n-grams, are easily constructed from text corpora irrespective of the target language.

Incorporating Part-of-Speech (PoS) based statistics provides additional precision. Also, user style and preferences can be accounted for by maintaining

n-grams collected from user texts. Thus, the combination of different n-gram statistics constitutes the base of the FASTY predictor providing a baseline performance for all target languages. Other modules interact with these results and improve on them.

Abbreviation Expansion Abbreviation-expansion is a technique in which a combination of characters, an "abbreviation," is used to represent a word, phrase or command sequence. When an abbreviation is typed, it expands to the assigned word, phrase or command sequence. Abbreviation expansion is smoothly integrated into the ordinary prediction process, i.e. if the user types the beginning of an abbreviation the system is able to predict the abbreviation like an ordinary word (or phrase). In case of predicting the abbreviation code, the user interface can choose to show not only the completed abbreviation code but also the full expansion of the abbreviation in the prediction window.

Morphological Processing and Backup Lexicon The morphology component is not a prediction component per se, it rather performs auxiliary functions for other components. As a prediction resource it is used in contexts where the other components run out of predictions (e.g., if the correct (in the current context) inflected form of the word to be predicted is not contained in the n-gram tables it can be generated using the morphological lexicon).

Since one of FASTY's goals is to be able to suggest only wordforms appropriate for the current context, it is required that the system is able to perform morphological analysis and synthesis, and to extract the morphosyntactic features needed by the components dealing with checking syntactic appropriateness. Also compound prediction needs the morphosyntactic information of the compound parts to be able to correctly predict the linking elements.

Last but not least, if the frequency based lexica run out of words with a given prefix, the morphological lexicon–provided it is big enough–will serve as a "backup" lexicon and deliver additional solutions.

Morphological processing is be implemented via finite state-transducers, which provide very fast, bi-directional processing and allow for a very compact representation of huge lexica.

2.2 Grammar-based Prediction and Prediction Ranking

The primary purpose of the grammar-based module is to enhance the predictive power of FASTY and improve its precision by syntactic processing. Only predictions that are not in conflict with the grammar will be delivered. The component should also be able to fill in gaps in terms of missing forms in the prediction list delivered by the n-gram based prediction module. It is assumed that syntactically well motivated n-grams provide better predictions than those that are not.

The module is realised as a partial parser employing the UCP (Uppsala Chart Parser) formalism [8], [9]. It uses the same PoS tags as basic categories

as the core module. In contexts where all predictions by the core component are rejected on syntactic grounds (e.g., if the correctly inflected form of the word to be predicted is not contained in the n-gram tables) a prediction can be generated using a back-up morphological lexicon.

The predictive power of a syntactic processor lies in its ability to make predictions in terms of word classes (PoS) and inflectional forms. In doing so, it can handle larger contexts than PoS n-grams and the predictions should be safer. A syntactic processor is not, however, capable of predicting new lexical material. It has to be fed by input from the user or by predictions made by other modules of the system. Typically, the syntactic processor will analyse predictions made by the n-gram module. As a result of the analysis it will primarily categorise them into three categories

- predictions confirmed by the grammar
- predictions outside the scope of the grammar
- predictions turned down by the grammar

A syntactically based ranking of the members of the first and second types is foreseen. It will be based on frequency data on syntactic structures for the different languages. In the first version of the prediction system, the ranking of the members of the first two categories will follow the probabilistic ranking suggested by the n-gram module.

As a rule, the predictions turned down by the grammar will not be delivered to the controller. So, the user will not be annoyed with predictions that are syntactically impossible; further, if impossible forms are left out there will be more space left in the prediction list for other possible continuations and thus convergence to the continuation intended by the user is achieved earlier.

A refuted inflectional form may also indicate the need for filling a gap in the prediction list. In other words, the lemma of the refuted inflected form, may in fact be the one that the user is aiming for. In such a situation first the other predictions are examined if an admissible form of that lemma is already contained. If not, a syntactically appropriate form is searched for in the morphological back-up dictionary.

2.3 Compound Prediction

Nominal compounds (the most frequent type of neologisms) are formed as new words in three of the FASTY languages (i.e., Dutch, German and Swedish). The type frequency of compounds is quite high. Analysis of a German corpus showed that 47% of the wordforms are compounds, suggesting that they constitute a class of words that must be handled for a satisfactory coverage of the patterns of the language.

On the other hand, compounds tend to have a low token frequency, and they are often *hapax legomena*. This suggests that they are formed productively. As a consequence, even with the largest lexicon, many possible compounds will not be in the lexicon. Our corpus analysis also showed that the large majority of

compounds are made of words that independently occur in the corpus (and typically with higher frequency than that of the compounds they form). Therefore, it seems reasonable to conclude that compounds should mostly be predicted by a compound prediction device, rather than be stored in the lexicon. Given that both Dutch [2] and Swedish have also very productive compounding, similar arguments hold for these two languages.

Our data indicate that the large majority of German compounds is composed of two element compounds. Among those, the most common type is the one in which both the first and the second element are nouns (around 80% of the total in terms of both type and token frequency). Thus, our compound prediction module focuses on the prediction of the N+N structure. We will refer to the first/left element of compounds of this sort as the modifier, and to the second/right element as the head. This general structure is also valid for Dutch and Swedish.

The head prediction model Because of the low token frequency of compounds (vs. the high frequency of their components), compound prediction should be performed very conservatively. In our system, the compound prediction module will only be invoked if the user, after selecting a noun, types the backspace character to delete the space automatically inserted after the noun, providing an explicit signal that she intends to construct a compound.

That means, instead of trying to predict whole compounds, we limit ourselves to trying to predict the head of the compound, after the modifier has been identified. Since in this model the modifier is treated as an independent word, n-gram statistics will be collected counting left elements as independent words. In order to predict compound heads, the module will choose the candidate heads with the highest scores for the weighted sum of the following measures:

- Unigram probability of head
- Bigram probability of head in current context
- Tag-based trigram probability of head in current context
- Likelihood of head to occur in a compound
- Probability of the "compound-bigram" based on semantic classes

The measures proposed are based on very general distributional properties of (right-headed) compounds. They should be completely language independent. A more detailed account can be found in [1].

Treatment of linking suffixes In all three languages, a large portion of the N+N compound lemmas contains a linking suffix attached to the modifier. Our analysis showed that most modifier + linking suffix sequences are identical to inflectional forms of the modifier. The only pattern (in all three languages) that does not always correspond to an independently existing inflected form of the modifier are forms ending in the suffix *-s*.

In the head prediction model, modifier + linking suffix sequences identical to inflected forms of the modifier are handled implicitly, modifiers with the linking

suffix *-s* share a very limited number of endings. A study of the relevant patterns in each language leads to reliable heuristics enabling us to predict the distribution of the s-linker in a satisfactory way.

Interaction with other modules Predictions produced by the compound prediction module do not compete with predictions from other modules. Once the user invokes the compound prediction module, the module takes control of the output, and proposes completions for the current substring until the user makes a selection.

2.4 Collocation-based prediction

The n-gram model predicts words on the sole basis of their immediate context. However, long-distance relationships between a word and other words in the text in which it occurs can also be exploited to improve word prediction. In particular, since texts typically have a topic and are semantically coherent, the appearance in a text of a certain (content) word can be a cue that other, semantically related words, are likely to appear soon.

Trigger pairs A straightforward way to exploit this fact is the notion of a trigger pair, i.e. a pair of words A and B that are correlated in such way that the occurrence of the trigger word A earlier in a text will affect the probability of the occurrence of the triggered or target word B in the same text [7]. Of course, A and B can, and will often be, the same word (making the notion of recency promotion as used e.g. in [4] a special case of the more general trigger pair idea).

In order to construct trigger pairs, we need a measure of the "textual association strength" between words. Of course, the task of manually comparing all the words in a corpus and determining, for each pair, how "associated" the two members are, is not feasible, and it is also not clear that our association intuitions would be very reliable. Instead, various statistical association measures that can be automatically extracted from training corpora have been proposed and sometimes compared in the NLP literature (see for example [3], [6] and [7]). Most measures are based on the comparison of the number of times that a pair of words co-occur in the training text to the number of times that the two words would be predicted to occur if they were independent events. Clearly, the larger the deviation from this expected value, the more likely it is that the two words are textually correlated. We will first adopt the average mutual information between two words/events as our association strength measure, other measures (such as the ones listed in [5]) will be taken into account in the future.

Using trigger pairs during word prediction Each word in the dictionary is associated with a (possibly empty) list of its targets. With each target, the mutual information score of the relevant trigger pair is also stored. At run-time, for each recognised word the list of its targets (with their mutual information /

association strength score) is added to a table. If a target is already in the table, its score will be increased. During prediction, the collocation-based module will provide (additional) candidates (or promote words already in the prediction list) based on the scores of the words in the target table. An extension to stem-based trigger pairs will also be considered. However, in that case the collocation-based module will have to interact with the morphological component and the syntactic module, to make sure that, for each candidate target stem, only inflectional forms that are morpho-syntactically legal are generated.

3 Summary

We have described the language component of FASTY, a system designed to provide efficient Predictive Typing for European languages other than English. Building on standard technology a number of novel solutions have been developed to deal with problems like inflection and compounding. In particular, morphological processing and partial parsing have been integrated with standard statistical prediction, and a model for split-compound prediction has been introduced. The system is currently under development, first results are promising. A fully integrated prototype is to be expected end of 2003.

Acknowledgments

This work was supported by the European Union in the framework of the IST programme, project FASTY (IST-2000-25420). Financial support for ÖFAI is provided by the Austrian Federal Ministry of Education, Science and Culture.

References

1. Baroni M. Matiasek. J, Trost H.: Predicting the Components of German Nominal Compounds. In F. van Harmelen (ed.): *ECAI 2002. Proceedings of the 15th European Conference on Artificial Intelligence*, IOS Press, Amsterdam. (to appear)
2. Boij, G.: Compounding in Dutch, Rivista di Linguistica, **4**(1) (1992) 37–59
3. Brown P.F., Della Pietra V.J., DeSouza P.V., Lai J.C., Mercer R.L.: Class-based n-gram models of natural language. *Computational Linguistics* **18**(4) (1990) 467–479
4. Carlberger J.: Design and Implementation of a Probabilistic Word Prediction Program. Masters Thesis, Royal Institute of Technology (KTH). (1998)
5. Evert, S.: On lexical association measures, ms., IMS, Universitt Stuttgart, (2001)
6. Jelinek F.: Statistical Methods for Speech Recognition. MIT Press, Cambridge/Boston/London (1998)
7. Rosenfeld R.: A Maximum Entropy Approach to Adaptive Statistical Language Modeling. *Computer Speech and Language* **10** (1996) 187–228
8. Sgvall Hein, A.: A parser for Swedish. Status report for sve.ucp. Technical Report UCDL-R-83-2, Uppsala University. Center for Computational Linguistics (1983)
9. Weijnitz P.: Uppsala Chart Parser Light. System Documentation. In: Working Papers in Computational Linguistics & Language Engineering **12**. Department of Linguistics, Uppsala University (1999)