Dept. for Speech, Music and Hearing
Quarterly Progress and Status Report

Constructing a database for a new Word Prediction System

Carlberger, A. and Hunnicutt, S. and Carlberger, J. and Strömstedt, G. and Wachtmeister, H.

KTH Computer Science and Communication

journal: TMH-QPSR
volume: 37
number: 2
year: 1996
pages: 101-104

http://www.speech.kth.se/qpsr

Constructing a database for a new word prediction system

Alice Carlberger, Sheri Hunnicutt, Johan Carlberger, Gunnar Strömstedt & Henrik Wachtmeister
Department of Speech, Music & Hearing, KTH

Abstract
A strictly frequency-based adaptive lexical prediction program, used mainly by persons with motoric handicaps or linguistic disabilities such as mild aphasia and dyslexia, is undergoing major development to improve the prediction function. Such development includes the change to a probability-based system, scope extension, and addition of grammatical, phrasal, and semantic information to the database.

Introduction

Profet is a statistically based word prediction system that was developed at our laboratory and has been used for a number of years as a writing aid by persons with motoric disabilities and/or linguistic impairments such as dyslexia and mild aphasia (Hunnicutt, 1986, 1989). As soon as the user starts spelling a word, Profet displays a list of one to nine word alternatives, depending on the selected setting. The user may choose one of these words with the function keys or continue to type the next letter(s) of the word until either the target word appears in the word alternative list or the word has been spelled completely. Choosing words from the word alternative list can save the user up to 26% in keystrokes and 34% in letters typed, when only one word is predicted (Hunnicutt, 1986). Better results are obtained for more word choices.

The current prediction function uses three main information sources: the word frequency lexicon, the word pair lexicon, and the subject lexicon. The first of these contains approximately 10,000 words with frequency information. The word pair lexicon consists of some 3,000 reference words each of which has an associated list of one to nine words that frequently succeed it. The successors are arranged according to frequency within the paradigm. The purpose of the subject lexicon is to allow the prediction system to adapt to the user's language by adapting the word frequency lexicon with those words of the user that are not in the lexicon or that have a rank higher than 1000. Work on the new prediction system commenced in July 1995.

Goal

It is our goal to enhance the current prediction capability by extension of scope, addition of grammatical, phrasal, and semantic information, and abandonment of a strictly frequency-based system for a probability-based one, allowing information from multiple sources to be weighted appropriately for each prediction.

Improving prediction

Methodology

Extending the scope and adding grammatical information

Extension of scope and addition of grammatical information constitutes the main focus of the development project. In the current version of Profet, a word is predicted solely on account of the previously typed word (through look-up in the bigram lexicon), unless one or several of its letters have been typed, in which case prediction is based entirely on single word frequency (through look-up in the single word frequency lexicon) while the previous word is completely ignored. In the new version, however, the predictor scope will be extended, so that even information about words preceding the previous word will be used in the prediction. Prediction will also be based on previous word(s) *even* after typing any letters of the new word. The extended scope should lead to word suggestions that are grammatically more correct than those presently given.

Example:

Current version: Den andra	sidan
	stycket*
	länder*
	ord*
	håll*
New version: Den andra	sidan
	handen
	vägen
	personen
	gången

(* = ungrammatical)

The Swedish language contains a large percentage of homographs. For instance, a word like "basar" has three meanings: noun singular indefinite ('bazaar'), noun plural indefinite ('bases'), and verb present tense ('is/are/am in charge of'). Therefore, the addition of grammatical tags to database words is expected to enhance prediction capability.

Since the current database lacks grammatical information as well as statistics for the occurrence of sequences longer than two contiguous words, a new database must be built. Besides bigrams (word and grammatical tag pairs with co-occurrence statistics), we intend to add trigrams as well as collocations (non-contiguous sequential word and grammatical tag bigrams with 2-5 intervening words). An initial effort to extract syntactic collocations (subject-verb, verb-object, subject-object, preposition-noun, adjective-noun) from the two corpora described below was aborted due to the internal complexity of the sentences in the corpora. Instead, a method is being employed whereby the collocations are extracted without regard to syntactic roles. All information in the new database, including collocations, must be extracted from one single corpus in order to warrant the implementation of a probabilistic prediction function. Two corpora are currently available to us, each of which will now be presented briefly.

We have based our system of 128 grammatical tags on that of the Stockholm Umeå Corpus (SUC). (See Appendix A) The fine detail of this tag system necessitates a large corpus for the extraction of database information. While awaiting the forthcoming 1-million-word tagged SUC corpus, the 300,000-word tagged SUC subcorpus[1] is being used. We also have access to a 150-million-word corpus, a balanced untagged conglomerate of electronic texts (newspapers, government documents, novels, adolescent literature, legal and medical documents, and recipes)[2]. Our aim to construct a database with statistically congruent single and n-gram word and grammatical tag information, from which the new prediction function will be able to calculate the probabilities, presents us with the following options:

1. Extracting all database information from SUC.

Advantages: Statistical congruency; tag system consistency.

Disadvantage: Small corpus, which might lead to underrepresentation of certain tags and words.

2. Extracting all database information from conglomerate corpus.

Advantages: Statistical congruency; large amount of data, which should be more indicative than SUC of the "real" language.

Disadvantage: No grammatical information, which would necessitate the creation of a rudimentary tagger, since SUC tagger is not a public domain tool.

3. Extracting grammatical tags from SUC and word statistics from conglomerate corpus.

Advantages: Tagging more accurate than inventing rudimentary tagger and tagging acorpus automatically; word statistics based on large amount of data.

Disadvantages: Merging statistical data from different corpora incongruous; small corpus for tag extraction, which might lead to underrepresentation of certain tags. The word statistics in the two corpora could be checked against each other, but the relationship of the tags in the two corpora could not be established.

A pre-processing program has been developed by Jesper Högberg at our department. (See Appendix B.)

Adding lexicalized phrases

Currently, Profet presents one word on each line in the prediction window. The new version, however, will allow for the display of more than one word per line. This work was begun in the Fall of 1995. Our collaborator in this subproject, Jan Lindberg of the Department of Computational Linguistics at the University of Stockholm, has furnished us with approximately

[1] Currently available on CD-ROM through the European Corpus Initiative (ECI).

[2] Sources: Språkdata 24 million words, SRF Tal & Punkt 37 million words, Göteborgsposten 5 million words, and Pressens Bild 100 million words.

9,000 uninflected and 48,000 inflected lexicalized phrases. Jan's working definition of a "lexicalized phrase" is that it is a sequence of at least two, usually three or more, words that tend to co-occur so frequently that they are pronounced as one prosodic unit, e.g., "för all del", "tänka sig in i" or "så till den grad". For our purposes, the phonetic aspects are of no immediate interest to the Profet project itself. Rather, it is the capability to present a sequence of words in the prediction window as a cognitively coherent unit, since this would provide the user with yet another keysaving strategy. Here is an example:

Current version: I och	det
	jag
	att
	en
	den
New version: I och	för sig
	med
	med att
	utanför
	av sig självt

Adding semantic information

The work of adding semantic information was started in February 1996. While it is unlikely to result in any actual keystroke savings, it seems conducive to constructive thought in the writing process to present the user with suggestions of words that are semantically congruent with the preceding words. Therefore, we (in collaboration with Kenneth Nordberg, a student of computational linguistics at the University of Stockholm) are exploring the possibilities of furnishing the nouns, adjectives, and verbs in the database with semantic tags such as "human", "animate", and "inanimate". Nouns and adjectives would, of course, have one semantic tag. Verbs, on the other hand, would have from one to three semantic tags, depending upon their transitivity. The following example illustrates the advantage of a semantically sensitive prediction system:

Current version:	Det är en pratsam liten	stund*
		och*
		del*
		bit*
		flicka
New version:	Det är en pratsam liten	flicka
		pojke
		gubbe
		gumma
		farbror

* (= ungrammatical)

References

Hunnicutt S (1986). Lexical Prediction for a Text-to-Speech System. In: *Communication and Handicap: Aspects of Psychological Compensation and Technical Aids*, Hjelmquist E & Nilsson L-G, eds., Elsevier Science Publishers.

Hunnicutt S (1989). ACCESS: A Lexical Access Program, *Proc RESNA 12th Annual Conference*, June 25-30, New Orleans, 284-285.