Computer Detection of Typographical Errors

ROBERT MORRIS AND LORINDA L. CHERRY

Abstract—A computer program written for the UNIX time-sharing system reduces by several orders of magnitude the task of finding words in a document which contain typographical errors. The program is adaptive in the sense that it uses statistics from the document itself for its analysis.

In a first pass through the document, a table of digram and trigram frequencies is prepared. The second pass through the document breaks out individual words and compares the digrams and trigrams in each word with the frequencies from the table. An index is given to each word which reflects the hypothesis that the trigrams in the given word were produced from the same source that produced the trigram table. The words are sorted in decreasing order of their indices and printed. Printing is suppressed for words appearing in a table of 2726 common technical English words. The table is attached as Appendix B.

The author of a 108-page document needed less than ten minutes to scan the output and identify the misspelled words. There were a total of 30 misspelled words among the 386 words output and 23 of those occurred among the first 100 words output.

INTRODUCTION

A large portion of the usage of the UNIX time-sharing system at Murray Hill has been applied to the preparation of documents. A major part of the task of document production has been made automatic by the existing hardware and software. One of the most costly remaining parts of the overall job is the manual proofreading of the document in draft form.

The method described here substantially reduces the time required to perform part of this proofreading task, namely that of checking the correct spelling of individual words.

SYNOPSIS OF THE PROGRAM

A document is read by the program and an attempt is made to recognize the individual words in the text.

—all upper case characters are mapped into lower case,

—apostrophes are deleted from within words

—words broken across lines are reassembled

Two processes are then performed on the words as they are input. First they are scanned for their digram and trigram statistics and secondly they are sorted, duplicate words are cast out, and one copy of each word is stored in a temporary file.

Typical experiences with large files is that the number of distinct words remaining varies greatly depending on the author, but for a given author, the number of distinct words increases approximately as the square root of the total number of words. Typically a ten-thousand word document contains about 1500 distinct words.

Robert Morris and Lorinda L. Cherry, Bell Laboratories, Murray Hill, New Jersey.

Original manuscript received January 25, 1975.

As the words are being broken out, a count is kept of each kind of digram and trigram in the document. For example, if the word 'once' appears in the text, the counts corresponding to each of these five digrams and four trigrams are incremented.

.o on nc ce e.

.on onc nce ce.

where the '.' signifies the beginning or ending of a word. These statistics are retained for later processing.

In order that the statistics should give some meaningful results for very short documents, the digram and trigram tables are initialized with some statistics taken from typical technical English text. This has little effect on the results for large documents and can be suppressed by the user if desired.

The table of sorted words is compared with a table of 2726 common technical English words as used at Murray Hill and those words that appear in the table of common words are removed.

The table of common words contains the results of processing about one million words of technical text appearing in documents produced on the UNIX time-sharing system at Murray Hill. The words selected are those which have the highest probability of appearing in a given document. This is not the same as a list of the most frequently occurring words and, in fact, some rare words occur in the table. For example, the words 'murray' and 'abstract' are uncommon words in the documents sampled, yet they both appear once in virtually every document. A copy of this list of common words appears in Appendix A.

The number of words removed by consulting the table varies in practice from about 500 to about 1500 and is dependent more on the author than on the length of the document.

The remaining words from the original document are then read one by one a number is attached to each word by consulting the table of digrams and trigrams previously produced.

The number is an index of peculiarity and reflects the likelihood of the hypothesis that the trigrams in the given word were produced from the same source that produced the trigram table.

Each trigram in the current word is treated separately. An initial and terminal trigram is treated for each word so that the number of trigrams in a word becomes equal to the number of characters in the word. For each trigram T = (xyz) in the current word, the digram and trigram counts from the prepared table are accessed. Each such count is reduced by one to remove the effect of the current

word on the statistics. The resulting counts are n(xy), n(yz), and n(xyz). The index i(T) for the trigram T is set equal to

$$i(T) = \tfrac{1}{2}[\log n(xy) + \log n(yz)] - \log n(xyz)$$

This index is invariant with the size of the document; i.e., it needs no normalization. It measures the evidence against the hypothesis that the trigram was produced by the same source that produced the digram and trigram tables in the sense of References 1 and 2. In the case that one of the digram or trigram counts is zero, the log of that count is taken to be -10.

Several methods were tried to combine the trigram indices to obtain an index for the whole word including:

—the average of the trigram indices.
—the root-mean-square of the trigram indices, and
—the largest of the trigram indices.

All were moderately effective; the first tended to swamp out the effect of a really strange trigram in a long word and the latter was insensitive to a sequence of strange trigrams whose cumulative evidence should have made a word suspicious. The second appeared on trial to be a satisfactory compromise and was adopted.

The words with their attached indices are sorted on the index in decreasing order and printed in a three-column format. The indices have been normalized in such a way that a word with an index greater than 10 contains trigrams that are not representative of the remainder of the document. Such a word almost certainly appears just once in the document.

The results for a 108-page document typed into the UNIX system are representative of the results for large documents. The total number of words in this document was 19917. The number of distinct words was 1207 and of these, 821 were removed by reference to the list of common words, leaving the 386 words in Appendix B to be output. The total running time of the program for this document was about 3 minutes on a PDP 11-45 minicomputer.

The sample document uses a smaller-than-average vocabulary for technical writing because it is detailed and unusually repetitive description of a limited subject area. On the other hand, it is typical of documents of its size and larger that are typed into the UNIX system.

An author of the document took less than 10 minutes to scan the output and identify the misspelled words. A total of 30 misspelled words was found. The value of the trigram analysis is indicated by the fact that 23 of the misspelled words occurred among the first 100 words output and only 3 occurred among the last 100 words output.

Proofreading the original document and finding this many typographical errors would be a nearly impossible task. Experiments with other documents have shown that an author finds barely half of the misspelled words in his own document if the document is more than about 20 pages long, but that he finds virtually every misspelling by looking once through the output of this program.

Two keys to the high rate of error locating are (1) that the output from the program is exceedingly short and every word can be looked at rather carefully in only a few minutes, and (2) that the percentage of errors is sufficiently high, especially at the beginning of the output, to made the effort psychologically rewarding. The author, seeing some errors among the first few words, is induced to continue scanning the list with considerable care.

The misspelled words located by the author of the sample document have been marked by hand in Appendix B.

Considerable reduction in the volume of output could be achieved by discarding words that occurred in the original document more than once and retaining for further processing only those which occurred just once in the document. This extension met with considerable customer resistance since it concealed consistent misspellings. On the other hand, no attention is now paid to the potential value of treating specially words which occur just once. They evidently would repay some special attention.

The program does not by any means do the whole job of proofreading, and the draft document still has to be read, but with considerably less care. The program does not find missing or extra words or semantic nonsense.

The principal annoyance in using the program is a consequence of the initial word breakout which maps all letters into lower case and removes apostrophes and the like. The result of this processing and the (necessary) lack of any information about where the offending word occurred in the original makes it rather difficult to find an occasion, even with the use of a context editor.

TRIGRAM ANALYSIS

Early attempts to separate misspelled words from corrected English words by the use of preset trigram tables failed completely. It appeared that the trigram frequencies of a document, no matter how long, did not approach the trigram frequencies of a document by some other author or on some other subject.

A table of trigram frequencies was prepared using the 2006 most frequent words in the English language according to Reference 3 and this table was compared against actual samples of technical and non-technical English text. The only authors whose writings showed a useful correlation were Edmund Burke and Thomas Jefferson. In every other case (including, for example, Mark Twain) the prepared table was absolutely useless.

On the other hand, in any particular document, the trigram frequencies obtained from one part of the

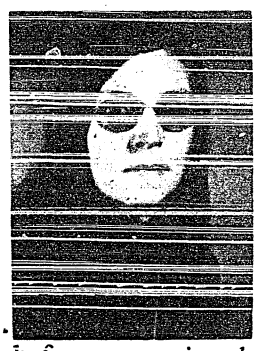
document were highly correlated with frequencies from the remaining parts of the document, particularly in technical writing. For this reason, the trigram frequencies were gathered from the document being analyzed even though this nearly doubled the running time of the program.

[1] Good, I. J. *Probability and the Weighing of Evidence.* Charles Griffin & Co., London, 1950.

[2] Kullback, S. and Leibler, R. A. On Information and Sufficiency. *Annals of Math. Stat.*, **22** (1951), pp. 79-86.

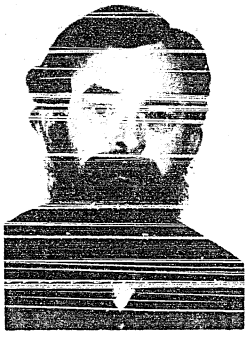
[3] Kučera, H. and Francis, W. *Computational Analysis of Present-Day American English.* Brown Univ. Press, Providence, 1967.

Contributors



LORINDA L. CHERRY and Robert Morris are members of the computing techniques research department in the Research, Communications Principles Division at Bell Laboratories' Murray Hill, N.J., location.

Ms. Cherry joined Bell Labs in 1966, and worked in the speech and acoustics area, and in the SAFEGUARD development area, before assuming her present position. She holds a B.A. in mathematics from the University of Delaware and the M.S. in computer science from the Stevens Institute of Technology.



MR. MORRIS has been conducting research in computer techniques since joining Bell Labs in 1960. He holds an A.B. in chemistry and an A.M. in mathematics, both from Harvard University. He is the author of a dozen articles and the recipient of two patents.