A NOTE ON UNDETECTED TYPING ERRORS

Although human proofreading is still necessary, small, topic-specific word lists in spelling programs will minimize the occurrence of undetected typing errors.

JAMES L. PETERSON

Computer programs to detect and correct spelling and typographic errors are fairly well understood and becoming quite common [8]. The interesting problems now deal mainly with the details of the design: the data structures and user interface, and not with the basic algorithms. Work is already proceeding on the construction of more advanced techniques for detecting syntactic and semantic errors [3, 6].

Despite the obvious value of spelling checkers, there can be problems. Perhaps the most important problems are mistakes made by the spelling checker. As with all pattern-recognition algorithms, mistakes can be of two kinds: (1) failure to accept a correctly spelled word, and (2) failure to reject an incorrectly spelled word. How bad are these problems?

Let us assume that the spelling checker is based on a lookup algorithm: The spelling checker maintains a word list of correctly spelled words. To check a word, the speller simply searches this list. If the word is found, it is assumed to be correctly spelled, and hence accepted; if it is not found, it is assumed to be incorrectly spelled, and hence rejected.

Clearly, with this algorithm, the speller will fail to accept correctly spelled words only if they are not on its list of approved words. To reduce the probability of this mistake, we can simply add more words to the approved word list, increasing its size. Ignoring the problems of storage and search time for large word lists, this approach may increase the probability of the second kind of mistake: failure to reject an incorrectly spelled word.

How can an incorrectly spelled word be mistak-
only recognized as correct by a spelling checker? Assume that the author/user meant to use the word \( x \), but mistyped the word as \( y \). (For example, house may be mistyped as horse.) This error will be undetected only if \( y \) is a correctly spelled word. It is admittedly not the desired word, but it is still a word on the approved word list. This kind of error can be detected only by more complex algorithms using syntactic or semantic information.

The probability of this kind of error—misspelling or mistyping a word as another word—should increase as the size of the word list increases. As the size of the word list increases, more obscure and unusual words will be added to the list. These words will no longer be detected as errors, but will be thought correct by the spelling checker. This article reports on research to determine the probability that a word can be mistyped as another word, as a function of the size of the word list.

**TYPES OF ERRORS**

First, we must redefine the problem. In our previous research into spelling checkers [9], we were unable to come up with any good information about how people misspell, so we have concentrated instead on how words are mistyped. The classical definition of typing errors by Damerau [4] indicates that 80 percent of typing errors are caused by

1. transposition of two adjacent letters,
2. one extra letter,
3. one missing letter, or
4. one wrong letter.

Our own studies have shown similar results. We found 360 errors in a computer copy of Webster's Seventh New Collegiate Dictionary [11] that had escaped detection since the file was keyboarded 15 years ago (including 10 errors in the original printed version). We also found 155 errors made by college students when they retyped the word division list of the U.S. Government Printing Office [5]. These errors were distributed as follows:

<table>
<thead>
<tr>
<th>Type of Error</th>
<th>GPO</th>
<th>Web 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transposition</td>
<td>4 (2.6%)</td>
<td>47 (13.1%)</td>
</tr>
<tr>
<td>One extra letter</td>
<td>29 (18.7%)</td>
<td>73 (20.3%)</td>
</tr>
<tr>
<td>One missing letter</td>
<td>49 (31.6%)</td>
<td>124 (34.4%)</td>
</tr>
<tr>
<td>One wrong letter</td>
<td>62 (40.0%)</td>
<td>97 (26.9%)</td>
</tr>
</tbody>
</table>

**Total** 144 (92.9%) 341 (94.7%)

The next most common errors were two extra letters, two missing letters, or two letters transposed around a third (such as procedure instead of procedure).

Although these results are similar, they are not sufficiently accurate or stable to provide good estimates of the probabilities of different kinds of errors or of key-level errors. There is almost certainly an error distribution for each key, due in large part to keyboard layout; a \( t \) is more likely to be mistyped as an \( r \) as a \( q \), while a \( q \) is more likely to be mistyped as an \( a \) than an \( r \). We have assumed simply that these four types of errors are all equally probable.

**WORD LISTS**

Over the last few years, we have collected some 16 word lists from various sources, including

- *Webster's Seventh New Collegiate Dictionary*, derived from the work of Olney at SDC and Alberga at IBM Yorktown [10] (81,627 words);
- *Webster's Second International Dictionary*, obtained over the ARPAnet, apparently keyboarded by students (234,933 words);
- *Longman's Dictionary of Contemporary English*, a British dictionary; the computer files were derived from the printer's phototypesetting tapes (20,506 words);
- *Random House Dictionary*, used in the Random House Proofreader, a spelling program from Aspen Software Company (50,030 words);
- the Proof spelling checker developed by IBM (70,399 words);
- the spell program for Unix® from Bell Labs (24,473 words);

All 16 word lists were merged to produce one master list of 369,546 words in a file of 3,799,425 bytes (an average of 9.3 characters plus a separator byte per word).

**THE EXPERIMENT**

A program was written to read the master word list and generate each of the possible “mistyped” versions of the word, according to the four main typing errors. Only one occurrence of one error per mistyping was considered, and each mistyped word was then checked to see if it was some other word. It took five days on a dedicated VAX 11/780—some 158,583.0 seconds of CPU time (about 44 hours)—to execute this program.

The result was a file of 15,609,965 bytes with 988,192 entries. Each entry consisted of two words and a code describing how one could be mistyped to

Unix is a trademark of AT&T Bell Laboratories.
be the other. Of the 369,546 unique words in our master list, 153,664 of them cannot be mistyped as some other word; 215,882 can be mistyped as some other word. In the extreme, for example, the word sat can be mistyped as 54 other words.

Of the 988,192 possible words that can be mistyped as another word,

616,210 were due to one wrong letter,
180,559 were due to one extra letter,
180,559 were due to one missing letter, and
10,864 were due to transposing two letters.

The numbers for one extra and one missing letter are the same because, if word x can be mistyped as word y by adding one extra letter, then y can be mistyped as x by deleting that extra letter. Similarly, the number of errors for wrong letters and transpositions should be even.

How many typing errors are possible? Assuming an alphabet of 28 characters (the 26 letters plus the hyphen and the apostrophe) and a word with n letters, the four typing errors could cause the following numbers of errors:

One wrong letter $27n$
One extra letter $28(n + 1)$
One missing letter $n$
Transposition $n - 1$

Thus, $27n + 27$ typing errors are possible for a word of length n. This allows, for our master list of 369,546 words, 205,480,845 possible mistyped words. Since only 988,192 of these are words, the probability of mistyping a word as another word is $\frac{988,192}{205,480,845}$ or about half a percent, for our full list.

VARYING THE SIZE OF THE WORD LIST

Of course, no one uses our entire word list in a spelling checker; we need to repeat this study for smaller word lists. There are, however, $2^{369,546}$ sublists for a word list of 369,546 words. It would be impossible (and pretty useless) to do this for every such sublist. We assumed instead that a word list of length m would be the m most-common words. Then any word on a list of size m would also be on each list of size $m' > m$. This allows us to determine the words in a word list as its size varies, if we know the frequency of the words.

There have been at least three major word frequency studies: the Brown Corpus [7] (50,406 unique words from a sample of 1,014,232), the American Heritage Intermediate Corpus [2] (66,741 unique words from a sample of 5,088,721), and a study of word frequency in the New York Times wire service stories [1] (70,144 unique words from a sample of 833,155). Although these studies provided some frequency information, they were not broad enough to provide complete information; only 22,439 words were on all three studies. To produce complete frequency information, we averaged the frequencies of the three major studies. Words that were in none of the studies were assigned a frequency corresponding to the number of times they occurred in our set of word lists. This produced a frequency for each word ranging from 1 to 323,302,702. Summing these frequencies and dividing by the sum give probabilities (ranging from 0.052 to 0.00000000016).

Sorting by frequency allows us to assign ranks to the words. The most frequent word is the, with of, and, to, a, in, is, that, it, and he completing the top 10. Words that have the same frequency (to the accuracy available from the existing studies) are all assigned the same rank. Thus, the 27,943 words that occurred in only one word list were all assigned the same rank: 369,546.

Now assume that a word x of rank p can be mistyped as a word y of rank q. If $p < q$ (i.e., x is more frequent than y), then the mistake will be detected for all lists with only words of rank less than $q$. If the word list expands to $q$ or more words, the mistake will go undetected. In general, the mistake will be undetected if the length of the word list is greater than or equal to $\max(p, q)$.

For example, the word house is rank 139; horse is rank 577. Thus, if house is mistyped as horse, it will be undetected if the length of the word list is greater than or equal to 577. If horse is mistyped as house, it will also be an undetected error for a word list of more than 577 words. (In fact, this error would be undetected for word lists from 139 to 577 words also, but in this case horse would not be considered a valid word as it is not on the word list. We have not considered the problem of mistyping a "nonword" as a word.)

RESULTS

Figure 1 (p. 636) shows the number of undetected errors as the size of the word list varies, from 1 to 369,546 words. The number of undetected errors appears to increase linearly with the size of the word list. Clearly, however, the number of errors that are undetected will almost always increase as the size of the word list increases; undetected errors remain undetected as more words are added to the word list. We should look, however, not only at the total number of undetected errors, but at the fraction of all errors that are undetected as well. As the word list grows, the probability that a given mistyping
would go undetected increases, but of course, with more words on the word list, there are more mistypings possible. Figure 2 shows the fraction of undetected typing errors as the size of the word list varies. For a word list of size $m$, this is the number of undetected typing errors for $m$ (from Figure 1) divided by the sum, over all $m$ words in the list, of the number of possible typing errors ($57n + 27$ for each word of length $n$). Although the fraction of undetected errors does increase as the word list grows to 50,000 or 60,000 words, it then levels off and is fairly stable, over the long run, at about 0.5 percent.

It would appear, then, that only 1 word out of every 200 mistyped words would accidentally become another word, escaping detection by a spelling checker. This result, however, ignores two important observations. First, frequent words tend to be shorter than less frequent words. And, second, short words, if mistyped, are more likely to be undetected than longer words, since proportionally more short words are on a word list. Of the 784 ($28 \times 28$) possible two-letter words, 431 (55 percent) are valid (in the complete 369,546 word list). On the other hand, only 21,611 words of 10 characters (out of a potential 296,196,766,695,424) are valid.

Figure 3 reinforces the conclusion that most undetected typing errors come from the short frequent words that are in every word list by measuring the probability of an undetected typing error in running text. Since small words are both more frequent and more likely to produce undetected typing errors, the probability of an undetected typing error varies directly with the size of the word list, ranging from 2 percent to almost 16 percent of all typing errors.

It is important to note, however, that these numbers are only rough indicators of the actual problem. Although we used the best information available to us, the following questions are not well answered:

- What is the frequency of words in English? Our averaging of the three major word frequency studies indicates that some words may have a frequency as low as one occurrence out of 6,000,000,000 words. Accordingly, an accurate
programs should be kept small; a large word list is not necessarily a better word list. In particular, word lists used for spelling should be tailored for the particular author and topic for which it is to be used. Word lists used for checking computer science papers should generally not include medical, legal, and geographic words, for example.

We also see a need for more intelligent checking programs, such as syntax and semantic checkers.

A simple spelling checker cannot detect a word that is mistyped as another word. More complex approaches (using part of speech information, for example) may be able to detect such errors.

Finally, although programs can catch many spelling and typing problems, human proofreading is still necessary to detect and correct many errors.

CONCLUSIONS

There is a real danger that longer word lists will result in a substantial number of undetected typing errors. Almost one typing error out of six may be undetected. Therefore, word lists used in spelling study may require billions of words to be analyzed.

• What is the frequency of the four main typing errors?
• What is the distribution of errors caused by keyboard layout?
• How are words misspelled (rather than simply mistyped)?

In addition, we would expect all of these numbers to vary with the author/typist.

Despite these unanswered questions and their effect on the actual probability of an undetected spelling error, it would seem reasonable to conclude that a significant number of spelling and typing errors may be undetectable by a spelling program, particularly with large word lists.

REFERENCES