Information Processing & Management Vol. 27, No. 5, pp. 517–522, 1991
Printed in Great Britain.

0306-4573/91 $3.00 + .00

CONTEXT BASED SPELLING CORRECTION

ERIC MAYS, FRED J. DAMERAU, and ROBERT L. MERCER
IBM Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598, U.S.A.

(*Received* 6 *June* 1990; *accepted in final form* 3 *December* 1990)

Abstract—Some mistakes in spelling and typing produce correct words, such as typing "fig" when "fog" was intended. These errors are undetectable by traditional spelling correction techniques. In this paper we present a statistical technique capable of detecting and correcting some of these errors when they occur in sentences. Experimental results show that this technique is capable of detecting 76% of simple spelling errors and correcting 73%.

1. INTRODUCTION

A traditional spelling error detection program determines that a word is misspelled only when the word does not appear in the program's dictionary of known words. That is, if a word is in the dictionary it is determined to be correctly spelled, and conversely, if a word is not in the dictionary it is noted as potentially misspelled. Since some misspellings (or typing errors) result in a word contained in the dictionary, there are errors that are undetectable using the traditional technique. For example, when attempting to detect errors in the sentence *I saw the man it the park.*, the error produced by transforming *in* into *it* is undetectable.

However, some of these errors are detectable by employing syntactic, semantic, or pragmatic knowledge about language. In the previous example, syntactic knowledge could be employed to determine that the sentence is syntactically ill-formed, and thus may potentially contain an error. In addition to performing spelling error detection, many programs also assist in spelling error correction by offering a set of candidate corrections that are "close" to the misspelled word. Here also, linguistic knowledge could be used to suggest likely corrections. Continuing the example, syntactic knowledge determines that substituting *in* for *it* produces a well-formed sentence, whereas substituting *is* does not.

The task underlying context based spelling correction is determining the relative degree of syntactic or semantic well-formedness among alternative sentences. The question is in what form linguistic knowledge should be represented and utilized to aid in this determination. For example, one possible form for representing syntactic knowledge is as a grammar. A parser could then determine which of the alternative sentences are those accepted by the grammar. There is a vast amount of knowledge that has been gathered by linguists and applied in computational settings. Nevertheless, we judge that applying linguistic knowledge in this form is a substantial undertaking.

We have studied the effectiveness of applying a statistical model of language that has demonstrated success in the task of speech recognition [1]. In this model, syntactic, semantic, and pragmatic knowledge is conflated into word trigram conditional probabilities. These conditional probabilities are derived from statistics gathered from large bodies of text.

Using the traditional approach to spelling error detection, a tension exists between the size of the dictionary and the number of undetectable spelling errors. As the number of words in the dictionary increases so does the number of undetectable spelling errors. For example, in the sentence, *I submit chat is what is happening in this case.*, the misspelling of *that* as *chat* would be detected if *chat* were not in the dictionary. On the other hand, if *chat* were in the dictionary the misspelling would not be detected. In [2] it was found that the (frequency weighted) number of potential undetected typing errors as a percentage of all typing errors ranges from 10% for a 50,000 word dictionary to 15% for a 350,000 word

dictionary. In one study [3] of a 20 million word corpus, utilizing a 60,000 word dictionary resulted in 50 times fewer errors than a 50,000 word dictionary, when comparing the ratio of undetected typing errors to incorrectly determined misspellings.

Note that traditional spelling correctors attempt to reduce this phenomenon by varying the words in the dictionary by subject area such as medicine or law. Additionally, certain words in the dictionary may be noted as uncommon, in which case the program may note this fact to the user. These techniques do not, however, take into account the specific context of word usage in a sentence or paragraph.

For the purposes of the experiment we describe here, as well as the experiments reported in [3] and [2], a spelling error is more precisely a typing error that is obtained by exactly one of the following four transformations:

- Adding an extra letter (e.g., mistyping *the* as *thea*)
- Deleting a letter (e.g., *the* as *th*)
- Replacing a letter (e.g., *the* as *ahe*)
- Transposing two adjacent letters (e.g., *the* as *hte*)

Errors of this kind typically account for over 80% of all errors [4,2], although special circumstances may show different statistics [5].

The remainder of this paper is organized as follows. In the next section, we develop the spelling correction process model and the language model. Section 3 describes the experiment we performed to evaluate the effectiveness of this approach. A detailed example is given in section 4. We conclude with a few brief observations.

2. LANGUAGE MODEL

We model spelling correction as a process similar to the model used in speech recognition [1]. See Fig. 1.

The text generator produces a word string w. As a result of a transformation performed by the speller and typist a word string y is produced which may differ from w. The linguistic decoder is to determine the word string $\hat{w}$ based on y. That is, choose $\hat{w}$ such that it is the w which maximizes the conditional probability of w given y, $P(w|y)$. Now,

$$P(\hat{w}|y) = \max_{w} P(w|y).$$

Using Bayes' formula,

$$P(w|y) = \frac{P(w)P(y|w)}{P(y)}$$

and since $P(y)$ does not depend on w, we have

$$P(\hat{w}|y) = \max_{w} P(w)P(y|w).$$

Here, $P(w)$ is the probability that the word string w will be produced by the text generator, and $P(y|w)$ is the probability that the speller and typist will transform the word string w into the word string y. $P(w)$ may be approximated using the word trigram model. The word trigram model utilizes the conditional probability of a word given two prior

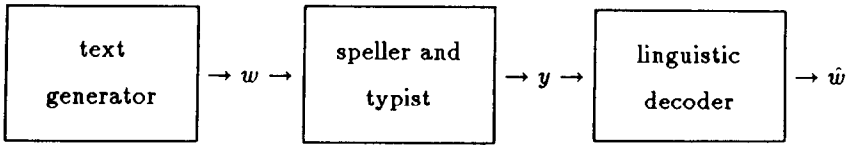

Fig. 1. Spelling correction model.

words. These probabilities are derived from three-word sequence statistics extracted from a large text sample. The probability of a sentence w_1, w_2, w_3, w_4 is the product of $P(w_1)$, $P(w_2|w_1)$, $P(w_3|w_1w_2)$ and $P(w_4|w_2w_3)$. (By contrast, in a bigram model the probability for w_1, w_2, w_3, w_4 is the product of $P(w_1)$, $P(w_2|w_1)$, $P(w_3|w_2)$, and $P(w_4|w_3)$.) So for a word string w consisting of the words $w_1, \ldots, w_n$,

$$P(w) = \prod_{i=1}^{n} P(w_n|w_{n-2}w_{n-1}).$$

We make the assumption that the transformations performed by the speller and typist do not insert or delete words and do not depend on adjacent words. Thus,

$$P(y|w) = \prod_{i=1}^{n} P(y_i|w_i).$$

The word string y is observed. From y, the set of possible word strings w is determined. There are many possible models of errors introduced by the speller and typist. If a speller or typist could produce y_i when w_i is intended, then y_i is in the *confusion set* w_i^c. Note that w_i is in w_i^c. For example, the confusion set for a word might include all simple misspellings of the word. In the case where the confusion set is determined by applying exactly one of the four basic transformations described earlier, we assume each element other than the word itself is equally likely, thus (where α is a constant),

$$P(y_i|w_i) = \begin{cases} \alpha & \text{if } y_i = w_i \\ (1-\alpha)/(|w_i^c| - 1) & \text{if } y_i \neq w_i \end{cases}.$$

That is, if the hypothesized ith word in the sentence, w_i, is identical to the observed ith word in the sentence, y_i, then $P(y_i|w_i)$ is the constant α. When w_i differs from y_i the remaining probability, $(1-\alpha)$, is distributed equally over the other elements in the confusion set, w_i^c. One can think of α as the a priori belief that the observed input word is correct.

The trigram probabilities are estimated by gathering statistics from large bodies of text using the methods described in [1]. This model is used to judge the relative well-formedness of sentences. That is, given two sentences s_1 and s_2, if s_1 is more well-formed than s_2, then $P(s_1) > P(s_2)$. The performance of the language model is determined by the degree to which this relationship between well-formedness and probability is achieved.

Note that α is a constant which must be determined by experimentation. A result of our experiment is an initial determination for a reasonable value of α. This is obtained by varying α and measuring the performance of the technique. However, this determination does not result in a provably optimal value.

3. EXPERIMENT

We performed an experiment to assess the viability of this technique. A 20,000 word vocabulary and associated trigram probabilities from the IBM speech recognition project [1] were used. One hundred sentences were randomly selected from the AP newswire and transcripts of the Canadian Parliament (50 from each), where the sentences consisted solely of words in our vocabulary. Neither the AP newswire nor the Canadian Parliament transcripts were part of the text sample from which the trigram probabilities were derived. Each of these sentences was assumed to have no errors. By systematically generating misspellings from the 100 sentences, it is possible to measure the performance of this technique, based on whether or not the misspelled sentence was corrected to the original sentence. The misspelled sentences were generated by substituting all of the misspellings for all of the words exactly once. Note that by a misspelling here, we mean an application of one of the basic transformations that resulted in another word in our vocabulary. Additionally, we are not concerned with misspellings of a word outside of the vocabulary that result in a word contained in the vocabulary. The set of 100 sentences produced 8628 misspelled sen-

tences. Additionally, we want to measure the performance of the technique on the original sentence, since transforming an original sentence into some other sentence introduces an error. In our implementation of the technique we were able to perform an exhaustive search, since pruning could be performed based on the use of trigrams, and the size of the confusion set is at most 30 for the misspellings (and vocabulary) we were considering.

Table 1 summarizes the results for varying values of α. The *original* column indicates the percentage of the original input sentences that were changed to some other sentence. The *changed* column indicates the percentage of the misspelled sentences changed to some other sentence. The *correct* column indicates the percentage of the *changed* misspelled sentences that were changed correctly. The *composite* column indicates the percentage of the misspelled sentences that were changed correctly (that is, the composite of *correct* and *changed*).

As α increases the percentage of misspelled sentences changed to some other sentence decreases while the correctness of the changes increases. While the composite correctness is maximum at $\alpha = .99$, note that three of the original 100 sentences were changed to some other sentence. At $\alpha = .999$ one of the original sentences was changed, while at $\alpha = .9999$ none of the original sentences were changed. From these results we conclude that a reasonable value for α lies in the range between 0.99 and 0.999. However, a larger set of sentences should be employed to evaluate the effect of α with respect to changing the original sentence into some other sentence. Ideally, α would maximize the composite correctness while minimizing the errors on the original sentences. For a suitably large set of sentences it may not be possible to achieve no errors on the original sentences. Thus one would need to express the desired optimum for α as maximizing a linear combination of composite correctness and original error rate.

We should note that some of the misspelled sequences we examined appeared to be plausible sentences that would not cause a human reader to make a correction without resorting to the context surrounding the sentence. For example, the sentence *I, of course, do not know what is in the document cabled by the minister.* is a misspelling of the sentence *I, of course, do not know what is in the document tabled by the minister.*, which our technique failed to correct. Also, the sentence *I, of course, do not know what is in the document tailed by the minister.* was changed incorrectly to the sentence *I, of course, do not know what is in the document mailed by the minister.* Our judgment is that these distinctions are nearly impossible to make outside of a larger context.

4. EXTENDED EXAMPLE

The following is an extended example of the technique and the experimental procedure. One of the sentences from the transcripts of the Canadian Parliament is sentence 1. By misspelling the first word in sentence 1, sentence 2 results. One of the misspelled sentences resulting from misspelling the second word in sentence 1 is sentence 3. The calculations in this example are based on the misspellings for words given in Table 2 and the trigram probabilities given in Table 3. The beginning of a sentence is noted by the token BOS. In our implementation, all calculations are based on (natural) log. Since only comparisons are relevant, this allows multiplications to be replaced by additions. The log probabilities of the sentences are: $\log P(S1) = -39.0574$, $\log P(S2) = -43.7417$, $\log P(S3) = -52.0891$.

Table 1. Experimental results

α	Original	Changed	Correct	Composite
.9000	15.0	94.4	78.7	74.4
.9900	3.0	86.9	90.9	79.0
.9990	1.0	76.7	95.4	73.2
.9999	0.0	63.7	97.0	61.8

Table 2. Words and misspellings

Word	Misspellings
I	a
submit	summit submits
a	I at as an am ad ab ya pa ha
summit	submit summits

Sentence 1. *I submit that is what is happening in this case.*
Sentence 2. *a submit that is what is happening in this case.*
Sentence 3. *I summit that is what is happening in this case.*

An attempt to correct sentence 1 will result in sentences 2 and 3 being considered as alternatives. By varying α the results in Table 4 are obtained. Notice that the conditional probability $\log P(S1|S1)$ differs from $\log P(S1)$ due to the contribution of α. That is, $\log P(S1|S1) = \log P(S1) + 11 \log \alpha$, since the sentence contains 11 trigrams. Similarly, $\log P(S2|S1) = \log P(S2) + 10 \log \alpha + \log((1 - \alpha)/10)$, and $\log P(S3|S1) = \log P(S1) + 10 \log \alpha + \log((1 - \alpha)/2)$. Note these last two cases differ due to the size of the confusion sets of *a* and *summit*. The value $\log P(S1|S1)$ is greater than $\log P(S2|S1)$ and $\log P(S3|S1)$ for all values of α considered in the experiment. Thus, for all experimental values of α the original sentence 1 is preferred over the alternatives, sentence 2 and sentence 3.

Suppose that sentence 2 were the input sentence. Then, the results in Table 5 obtain. Sentence 2 is preferred over the alternative sentence 1 except when $\alpha = .9$ and $\alpha = .99$. That is, the technique fails to correct sentence 2 to sentence 1 except when $\alpha = .9$ and $\alpha = .99$.

If sentence 3 were the input sentence, then the results in Table 6 obtain. In this case the original sentence 1 is preferred over the misspelled sentence 3 for all experimental values of α.

Table 3. Trigram log probabilities

w_1	w_2	w_3	$logP(w_3\|w_1w_2)$
BOS	BOS	I	-3.4763
BOS	I	submit	-8.4775
I	submit	that	-1.2305
submit	that	is	-4.7431
that	is	what	-3.0488
is	what	is	-3.0719
what	is	happening	-4.8898
is	happening	in	-1.7256
happening	in	this	-3.8423
in	this	case	-2.4928
this	case	.	-2.0586
BOS	BOS	a	-3.9681
BOS	a	submit	-10.2067
a	submit	that	-3.6938
BOS	I	summit	-18.4825
I	summit	that	-5.4944
summit	that	is	-3.5060

Table 4. Correcting sentence S1

α	$logP(S1\|S1)$	$logP(S2\|S1)$	$logP(S3\|S1)$
0.9000	-40.2164	-49.4005	-56.1385
0.9900	-39.1680	-50.7500	-57.4879
0.9990	-39.0684	-52.9621	-59.7000
0.9999	-39.0585	-55.2555	-61.9934

Table 5. Correcting sentence S2

α	$logP(S2\|S2)$	$logP(S1\|S2)$
0.9000	-44.9007	-42.4136
0.9900	-43.8523	-43.7631
0.9990	-43.7527	-45.9752
0.9999	-43.7428	-48.2686

Table 6. Correcting sentence S3

α	$logP(S3\|S3)$	$logP(S1\|S3)$
0.9000	-53.2481	-43.1067
0.9900	-52.1997	-44.4562
0.9990	-52.1001	-46.6683
0.9999	-52.0902	-48.9617

5. CONCLUSION

The experimental results presented indicate the initial success of applying maximum likelihood techniques to the problem of context based spelling correction. Beyond the direct application of this technique is the possibility of using results such as this as a benchmark against which the performance of computational linguistic models may be compared. Our intuition is that context based spelling correction is a task that people perform fairly well. In attempting to improve on the performance of computational linguistic systems, and to a certain extent theories as well, context based spelling correction is a well defined task against which progress can be measured.

REFERENCES

1. Bahl, L.R.; Jelinek, F.; Mercer, R.L. A maximum likelihood approach to continuous speech recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence *5*(2): 179–190; 1983.
2. Peterson, J.L. A note on undetected typing errors. Communications of the ACM *29*(7): 633–637; 1986.
3. Damerau, F.J.; Mays, E. An examination of undetected typing errors. Information Processing and Management *25*(6): 659–664; 1989.
4. Damerau, F.J. A technique for computer detection and correction of spelling errors. Communications of the ACM *7*(3): 171–176; 1964.
5. Mitton, R. Spelling checkers, spelling correctors, and the misspellings of poor spellers. Information Processing and Management *23*(5): 495–505; 1987.