

# TYPOGRAPHICAL AND ORTHOGRAPHICAL SPELLING ERROR CORRECTION

Kyongho Min\*, William H. Wilson\*, Yoo-Jin Moon†

\*School of Computer Science and Engineering  
The University of New South Wales  
Sydney NSW 2052, Australia  
{min,billw}@cse.unsw.edu.au

†Department of Management Information System  
Hannam University, Ojung-dong, Daeduk-ku, Daejeon, 300-791, Korea  
yjmoon@eve.hannam.ac.kr

## Abstract

This paper focuses on selection techniques for best correction of misspelt words at the lexical level. Spelling errors are introduced by either cognitive or typographical mistakes. A robust spelling correction algorithm is needed to cover both cognitive and typographical errors. For the most effective spelling correction system, various strategies are considered in this paper: ranking heuristics, correction algorithms, and correction priority strategies for the best selection. The strategies also take account of error types, syntactic information, word frequency statistics, and character distance. The findings show that it is very hard to generalise the spelling correction strategy for various types of data sets such as typographical, orthographical, and scanning errors.

## 1. Backgrounds

### 1.1. Typographical/Orthographical errors

Humans often make errors during communication, in either spoken or written language, at four levels: lexical, syntactic, semantic, and contextual (Eastman & McLean, 1981; Young, Eastman, & Oakman, 1991). This paper focuses on the correction of spelling errors such as typographical (Damerou, 1964; Pollock and Zamora, 1983), orthographical (Sterling 1983; Mitton, 1987), and scanning errors at the lexical level. The typographical errors involve regular forms of mistyping rather than cognitive errors (e.g. grammatical slips). For example, typographical errors may substitute a letter for a letter adjacent on the keyboard (Damerou, 1964; Pollock and Zamora, 1983). Orthographical errors arise from errors in cognitive processing such as guessing, or phonetic attempts at spelling, or selecting the wrong word - *to* for *too* or *two*, or *respectable* for *respective*.

Spelling errors have been studied in various contexts: typographical errors in a coordinated indexing and retrieval system (Damerou, 1964), misrecognition of written text (Galli and Yamada, 1968; Hanson, Riseman, and Fisher, 1976), spelling errors in scientific and scholarly text (Pollock and Zamora, 1983), and orthographical errors in spontaneous writings of children (Sterling, 1983; Mitton, 1987).

### 1.2. Approaches of Spelling Correction

Humans correct ungrammatical sentences by using preferred priority. Shanon (1973) tested the interpretation of spoken ungrammatical sentences in Hebrew, with one to three errors (e.g. violation of agreement rules in number, gender, and tense) in a sentence. Shanon found that humans preferred particular types of correction, for example in number-agreement violation, the verb is replaced rather than the noun, and in tense-agreement within a verb group.

Many researchers have worked on the correction of spelling errors, using various methods: dictionary look-up (Damerou, 1964), a code compression method (Pollock and Zamora, 1984); N-gram methods (Ullmann, 1977; Zamora, Pollock, and Zamora, 1981), a tagger (Atwell and Elliott, 1987), and integrated methods (Kese, Dudda, Heyer, & Kugler, 1992; Vosse, 1992; Carbonell and Hayes, 1983).

Different selection strategies for the best correction among candidate corrections were studied: word frequencies (Galli and Yamada, 1968), character distance (Min & Wilson 1995, 1999), frequency of the alternatives and prevalence of error types (Pollock and Zamora, 1984), and the probability of satisfying error rules (Yannakoudakis and Fawthrop, 1983).

## 2. Implementation of Two Spell Correctors

This section describes the design of two spelling correctors and two schemes for selection of the best correction. The spelling corrector employs a dictionary lookup technique, applied to two selection strategies: a character distance method and a word frequency method.

### 2.1. Implementation of two ranking heuristics

The spelling corrector generates re-spelling candidates made by  $55N + 26$  spelling transformations<sup>1</sup> which include the possibility that the first character of a word may be an error. After transforms generated by the spelling corrector are verified by dictionary lookup, a selection strategy is applied to the remaining verified words. The similarity between a misspelt word and a transform (i.e. a correct word) suggested is computed by two methods: word frequency statistics and character distance. The frequency statistics method depends on the fact that frequently used words are more prone to an error

---

<sup>1</sup> Comprising the  $26N$  possible substitutions, the  $27(N+1)$  possible additions, the  $N$  possible deletions, and the  $N-1$  possible transpositions, where  $N$  = the length of misspelt word, with 26 alphabetic characters plus apostrophe.

than words that are rarely used in a context (Peterson, 1980a, 1980b). The character distance method applies phenomena related to generation of typographical errors to the similarity measurement. In the case of frequency measurement, a more frequent word is considered more similar than less frequent word. However, in the character distance method, the possibility with the smallest character distance is chosen as the best correction.

• **Word Frequency Method**

For the purpose of selection of the best correction, the LOB-corpus (Atwell and Elliott, 1987) is used for word frequency information. For example, the misspelt word "whith" gave rise to the possibilities ("with" 7201), ("which" 4467), ("white" 261), and ("whit" 0) - where the numbers are the frequency of the associated words in the

corpus. The most frequent word among the possible corrections for "whith" (and hence the best correction according to this method) is thus "with". Note that "whit" is a correct word which does not occur in LOB corpus. If syntactic information is applied to the example above (e.g. ("whith" ADJ)), then just one possibility would be suggested (e.g. ("white" 261)). Thus using syntactic information increases the efficiency of a spelling corrector.

• **Character Distance Method**

This method uses a Pythagorean-type metric to measure the distance between a misspelt word and a possible correction, based on the Qwerty keyboard layout (Min and Wilson, 1995). The Qwerty keyboard is represented by two 2-dimensional arrays: one for the lower case keys and another for the upper case (shift) keys (Figure 1).

i\j	0	1	2	3	4	5	6	7	8	9	10	11
0	1	2	3	4	5	6	7	8	9	0	-	=
1	q	w	e	r	t	Y	u	i	o	p	[	]
2	a	s	d	f	g	H	j	k	l	;	'	
3	z	x	c	v	b	N	m	,	.	/		

Figure 1. Keyboard array (lower case) for Pythagorean metric.

For example, the character distance between *w* (3,5) and *n* (1,1) is  $\|(3,5)-(1,1)\| = \sqrt{2^2+4^2} \approx 4.47$ . Using this character distance, the scores for correction of "whith" are ("whit" 1.41), ("with" 2.24), ("which" 2.83), and ("white" 3.16). In contrast to the method using word frequencies, "whit" is the best correction. Once again, syntactic information could be used for filtering out more invalid spelling corrections, then "with" would be the best correction.

**2.2. Implementation of various experimental techniques**

We implmented 20 different spelling correctors and tested their performance at suggestion of the best correction among many alternatives:

1) four versions (called COMMON2.0/2.1 and CHAPTER2.0/2.1 in Table 1) are based on two factors -

*syntactic information* – either used or unused and *ranking heuristics* using either word frequency statistics or character distance;

2) eight versions (COMMON3.0, 3.1, 4.0, 4.1 and CHAPTER3.0, 3.1, 4.0, 4.1 in Table 1) differing in three factors - two factors as in 1) and *correction priority based on frequency of error types obtained from six test data sets*, e.g. correction priority based on general error type frequency or rarity (see table 3 in section 4.1); and

3) eight versions (COMMON3.01, 3.11, 4.01, 4.11 and CHAPTER3.01, 3.11, 4.01, 4.11 in Table 2) based on three factors - two factors from 1 and *correction priority based on frequency of specific error types from each test data*, e.g. correction priority based on the frequency each error type or on rarity of each test data.

version name	Syntactic Information used	Correction Priority Orders based on		Ranking alternatives based on
		Error frequency	Error rarity	
common2.0	no	-	-	word frequency
common2.1	yes	-	-	word frequency
common3.0	no	yes	no	word frequency
common3.1	no	no	yes	word frequency
common4.0	yes	yes	no	word frequency
common4.1	yes	no	yes	word frequency
chapter2.0	no	-	-	character distance
chapter2.1	yes	-	-	character distance
chapter3.0	no	yes	no	character distance
chapter3.1	no	no	yes	character distance
chapter4.0	yes	yes	no	character distance
chapter4.1	yes	no	yes	character distance

Table 1. Characteristics of 12 versions of spell correctors tested

version name	Syntactic Information used	Correction Priority Orders based on		Ranking alternatives based on
		Specific error frequency	Specific error rarity	
common3.01	no	yes	no	word frequency
common3.11	no	no	yes	word frequency
common4.01	yes	yes	no	word frequency
common4.11	yes	no	yes	word frequency
chapter3.01	no	yes	no	character distance
chapter3.11	no	no	yes	character distance
chapter4.01	yes	yes	no	character distance
chapter4.11	yes	no	yes	character distance

Table 2. Characteristics of 8 versions of spell correctors tested

Data name	total misspelt	Delete (%)	subst (%)	add (%)	transpose (%)	Multiple (%)
Appling1	192	52 (27.1)	36 (18.7)	34 (17.7)	7 (3.6)	63 (32.8)
Appling2	393	41 (10.4)	65 (16.5)	37 (9.4)	15 (3.8)	235 (59.8)
NED	96	26 (27.1)	18 (18.7)	19 (19.8)	26 (27.1)	7 (7.3)
Damerau	130	34 (26.2)	34 (26.2)	14 (10.8)	16 (12.3)	32 (24.6)
Email	399	145 (36.3)	46 (11.5)	91 (22.8)	52 (13.5)	65 (16.3)
Thesprev	33	15 (45.5)	7 (21.2)	1 (3.0)	1 (3.0)	9 (27.3)
Total (Average)	1243	313 (28.8%)	206 (18.8%)	196 (13.9%)	117 (10.5%)	411 (28.0%)

Table 3. Analysis of error types in terms of the number of misspellings

### 3. Experimental results for 20 versions of the spelling corrector

Table 4 shows tested results based on correction process priority, use of syntactic information, and two ranking heuristics. In addition, the table shows results of three rates that are computed as follows:

- Corrected rate = (total number of corrections (right or wrong) / (total number of tests));
- First hitting rate = (number of times that the correction ranked first was the same as the manual correction) / (total number of tests);
- False alarm rate = (total number of corrections – number of single errors) / (total number of corrections);

The versions of spelling corrector that used (manually generated) part-of-speech information reduced the false alarm rate<sup>2</sup> and the number of alternative corrections, and increased the first hitting rate.

Table 4 shows the results of effectiveness of syntactic information and correction priority for spelling correction. When using correction priority (as in spelling correctors with version numbers  $\geq 3.0$ ) to select the best correction among alternatives, the correction priority based on error frequency, in which the correction process for misspelt words with the most frequent error type is applied first, showed better results than that employing error rarity, by 9% to 10% in table 4.

Another finding was that the different ranking heuristics, for example between 'common3.0' and 'chapter3.0', made a small difference (2.1%) in the first hitting rate - see Table 4. Error correction methods using

character distance on the qwerty keyboard gave nearly the same results as that using word frequency statistics.

The spelling corrector cannot correct multiple spelling errors and so some words (with multiple errors) were corrected wrongly. In the case of false alarm rate, the use of syntactic information decreases the rate by 3% (34 misspellings) between COMMON3.0 and COMMON4.0 in Table 4.

The correction priority employing general error frequency (in versions 'common3.0' and 'chapter3.0') shows worse results by 0.5% in the first hitting rate than that employing specific error frequency (in versions 'common3.01' and 'chapter3.01'). However, the strategy employing correction priority based on general error rarity (in versions 'common3.1 & 4.1' and 'chapter3.1 & 4.1') shows better results by 1.4% to 2.5% in the first hitting rate than that employing correction priority based on specific error rarity (in versions 'common3.11 & 4.11' and 'chapter3.11 & 4.11'). In addition, the performance of spelling correction employing correction priority based on general error frequency showed better results than that based on specific error frequency (in versions 'common3.0 & 4.0' vs 'common3.01 & 4.01' and 'chapter3.0 & 4.0' vs 'chapter3.01 & 4.01').

Table 5 shows the degree of efficiency using syntactic information for spelling error correction. The versions with syntactic information reduced the complexity of selecting the first correction by around 12%. The versions with syntactic information reduced the number of suggested corrections that were in fact wrong. However, this rate would change as the size of the dictionary increased.

<sup>2</sup> This false alarm rate means that the corrections suggested do not include the manually corrected word because original word had multiple spelling errors.

Versions	Total test / single errors	corrected (%)	First hitting rate (%)	False alarm rate (%)
Common2.0	1243 / 835	979 (78.8%)	662 (53.3%)	147 (15.0%)
<b>Common2.1</b>	<b>1243 / 835</b>	<b>945 (76.0%)</b>	<b>734 (59.1%)</b>	<b>113 (12.0%)</b>
Common3.0	1243 / 835	979 (78.8%)	694 (55.8%)	147 (15.0%)
Common3.01	1243 / 835	979 (78.8%)	700 (56.3%)	147 (15.0%)
<b>Common4.0</b>	<b>1243 / 835</b>	<b>945 (76.0%)</b>	<b>769 (61.9%)</b>	<b>113 (12.0%)</b>
Common4.01	1243 / 835	945 (76.0%)	694 (55.8%)	113 (12.0%)
Common3.1	1243 / 835	979 (78.8%)	621 (50.0%)	147 (15.0%)
Common3.11	1243 / 835	979 (78.8%)	605 (48.7%)	147 (15.0%)
Common4.1	1243 / 835	945 (76.0%)	755 (60.7%)	113 (12.0%)
Common4.11	1243 / 835	945 (76.0%)	724 (58.2%)	113 (12.0%)
Chapter2.0	1243 / 835	979 (78.8%)	640 (51.5%)	147 (15.0%)
<b>Chapter2.1</b>	<b>1243 / 835</b>	<b>945 (76.0%)</b>	<b>712 (57.3%)</b>	<b>113 (12.0%)</b>
Chapter3.0	1243 / 835	979 (78.8%)	668 (53.7%)	147 (15.0%)
Chapter3.01	1243 / 835	979 (78.8%)	674 (54.2%)	147 (15.0%)
<b>Chapter4.0</b>	<b>1243 / 835</b>	<b>945 (76.0%)</b>	<b>767 (61.7%)</b>	<b>113 (12.0%)</b>
Chapter4.01	1243 / 835	945 (76.0%)	762 (61.3%)	113 (12.0%)
Chapter3.1	1243 / 835	979 (78.8%)	617 (49.6%)	147 (15.0%)
Chapter3.11	1243 / 835	979 (78.8%)	599 (48.2%)	147 (15.0%)
Chapter4.1	1243 / 835	945 (76.0%)	737 (59.3%)	113 (12.0%)
Chapter4.11	1243 / 835	945 (76.0%)	710 (57.1%)	113 (12.0%)

Table 4. Average Results of six data sets (1243 misspellings tested)

Versions	Total test	Single errors	Number of Single alternatives suggested	
			With syntactic information	Without syntactic Information
Appalling1	192	128	64	49
Appalling2	393	157	100	83
NED	96	89	70	47
Damerau	130	96	84	78
Email	399	340	275	238
Thesprev	33	25	16	14
Total (percentage)	-	835	609 / 835 (72.9%)	509 / 835 (61.0%)

Table 5. The number of single correction suggested (for the first corrections)

#### 4. Conclusion

This paper has focused on finding the best spelling correction based on various strategies, including ranking heuristics, various correction algorithms, and priority strategies by using error types, syntactic information, word frequency statistics and character distance. The findings are as follows:

(1) The COMMON-speller algorithm and the CHAPTER-speller algorithm differ by less than 2% in the first hitting rate. In the case of 'COMMON-speller', reliable word frequency statistics are needed for better results. However, the 'CHAPTER-speller' does not need such data and reflects the structure of a keyboard.

(2) The use of syntactic information for spelling correction increased the of the first hitting rate quite significantly, in fact by 6.1 – 10.7 %.

(3) The application of correction priority based on error type frequency increase the first hitting rate by 1.2 to 5.8% compared with that based on error type rarity.

(4) The employment of correction priority shows better results of the first hitting rate than the versions without the priority strategy by 2.8 - 4.4% (i.e. 'common2.1' vs 'common4.0' and 'chapter2.1' vs 'chapter4.0').

A spelling correction algorithm with higher-level information (e.g. syntactic and semantic information) would give better result for spelling correction. However, a generalised error correction algorithm is unlikely to apply with equal success to errors from different sources, as they are likely to have different characteristics.

#### 5. References

- Atwell, E. and Elliott, S., 1987. Dealing with Ill-formed English Text. In R. Garside, G. Leech, and G. Sampson (Eds.), *The Computational Analysis of English*. London, UK: Longman.
- Carbonell, J. and Hayes, P., 1983. Recovery Strategies for Parsing Extragrammatical Language. *American Journal of Computational Linguistics*, **9**(3-4): 123-146.
- Damerau, F., 1964. A Technique for Computer Detection and Correction of Spelling Errors. *Communications of the ACM*, **7**(3): 171-176.
- Eastman, C. and McLean, D., 1981. On the Need for Parsing Ill-formed Input. *American Journal of Computational Linguistics*, **7**(4): 257.
- Galli, E. and Yamada, H., 1968. Experimental Studies in Computer-Assisted Correction of Unorthographic Text. *IEEE Transactions on Engineering Writing and Speech*, **EWS-11**(2): 75-84.
- Hanson, A., Riseman, E., and Fisher, E., 1976. Context in Word Recognition. *Pattern Recognition*, **8**: 33-45.
- Kese, R., Dudda, F., Heyer, G., and Kugler, M., 1992. Extended Spelling Correction for German. *Third Conference on Applied Natural Language Processing*, 126-132.
- Min, K. and Wilson, W. H., 1995. Syntactic recovery and spelling correction of ill-formed sentences, *3rd Conference of the Australasian Cognitive Science (CogSci95)*, 82.
- Min, K. and Wilson, W. H., 1999. Syntactic recovery and spelling correction of ill-formed sentences, In J. Wiles and T. Dartnall (eds.), *Perspectives on Cognitive Science: Theories, Experiments, and Foundations*, vol.2, Stamford: Ablex.
- Mitton, R., 1987. Spelling Checkers, Spelling Correctors and the Misspellings of Poor Spellers. *Information Processing and Management*, **23**(5): 495-505.
- Peterson, J., 1980a. Computer Programs for Detecting and Correcting Spelling Errors. *Communications of the ACM*, **23**(12): 676-687.
- Peterson, J., 1980b. *Computer Programs for Spelling Correction: An Experiment in Program Design*. Berlin: Springer-Verlag.
- Pollock, J. and Zamora, A., 1983. Collection and Characterization of Spelling Errors in Scientific and Scholarly Text. *Journal of The American Society for Information Science*, **34**(1): 51-58.
- Pollock, J. and Zamora, A., 1984. Automatic Spelling Correction In Scientific and Scholarly Text. *Communications of the ACM*, **24**(4): 358-368.
- Shanon, B., 1973. Interpretation of Ungrammatical Sentences. *Journal of Verbal Learning and Verbal Behavior*, **12**: 389-400.
- Sterling, C., 1983. Spelling Errors in Context. *British Journal of Psychology*, **74**: 353-364.
- Ullmann, J., 1977. A Binary N-gram Technique for Automatic Correction of Substitution, Deletion, Insertion, and Reversal Errors in Words. *Computer Journal*, **20**(2): 141-147.
- Vosse, T., 1992. Detecting and Correcting Morpho-Syntactic Errors in Real Texts. *The Third Conference on Applied Natural Language Processing*, 111-118.
- Yannakoudakis, E. and Fawthrop, E., 1983. An Intelligent Spelling Error Detector. *Information Processing and Management*, **19**(2): 101-108.
- Young, C., Eastman, C., and Oakman, R., 1991. An Analysis of Ill-formed Input in Natural Language Queries to Document Retrieval Systems. *Information Processing and Management*, **27**(6): 615-622
- Zamora, E., Pollock, J., and Zamora, A., 1981. The Use of Trigram Analysis for Spelling Error Detection. *Information Processing and Management*, **17**(6): 305-316