An Evaluation of the Contextual Spelling Checker of Microsoft Office Word 2007

Graeme Hirst*
Department of Computer Science
University of Toronto
Toronto, Canada M5S 1A4

10 January 2008

Abstract

Microsoft Office Word 2007 includes a "contextual spelling checker" that is intended to find misspellings that nonetheless form correctly spelled words. In an evaluation on 1400 examples, it is found to have high precision but low recall — that is, it fails to find most errors, but when it does flag a possible error, it is almost always correct. However, its performance in terms of F is inferior to that of the trigrams-based method of Mays, Damerau, and Mercer (1991).

1 Real-word spelling correction

Most spelling checkers attempt to detect and correct only misspellings that result in a presumed *non-word* — a word that is not listed in the system's dictionary — and they therefore cannot deal with an error that just happens to form a real word in the dictionary, albeit not the word that the user intended. There has been much research in recent years on methods for detecting and correcting such *real-word errors* or *malapropisms* (we use the two terms interchangeably); for a review, see Hirst and Budanitsky (2005) and Wilcox-O'Hearn et al (2008).

The recently released Microsoft Office Word 2007 includes a "contextual spelling corrector" that attempts to detect and correct real-word errors (Microsoft 2006). A word that the system believes to be in error is flagged with a wavy blue underline, in contrast to Word's regular red underline for non-word errors, and suggested corrections are available in a pop-up menu or in the 'Spelling and Grammar' window. This system operates not only on content words but also closed-class words (e.g., *too* and *to*). It can detect cases where a word has been wrongly split into two (e.g., *through out* for *throughout*), and missing or spurious apostrophes (e.g., *corporations* for *corporation's*). It is not limited to a predefined set of frequently confounded words.

Here we report an evaluation of this system that was carried out in order to compare it with the word-trigram method of Mays *et al* (1991) and the lexical cohesion method of Hirst and Budanitsky (2005). (A detailed evaluation of the Mays *et al* method and a comparison with the lexical cohesion method is given by Wilcox-O'Hearn *et al* (2008).)

*This research was supported financially by the Natural Sciences and Engineering Research Council of Canada. I am grateful to Amber Wilcox-O'Hearn for comments and assistance.

2 Evaluation data

2.1 Basic data set

We used as test data the same data that Wilcox-O'Hearn *et al* used in their evaluation of Mays *et al*'s method, which in turn was a replication of the data used by Hirst and St-Onge (1998) and Hirst and Budanitsky (2005) to evaluate their methods.

The data consisted of 500 articles (approximately 300,000 words) from the 1987–89 *Wall Street Journal* corpus, with all headings, identifiers, and so on removed; that is, just a long stream of text. It is assumed that this data contains no errors — that is, that the *Wall Street Journal* contains no malapropisms or other typos. In fact, a few typos (both non-word and real-word) were noticed during the evaluation (see below), but they were small in number compared to the size of the text.

Malapropisms were randomly induced into this text at a frequency of approximately one word in 200. Specifically, any word whose base form was listed as a noun in WordNet (but regardless of whether it was used as a noun in the text; there was no syntactic analysis) was potentially replaced by any spelling variation found in the lexicon of the *ispell* spelling checker[1]. A *spelling variation* was defined as any word with an *edit distance* of 1 from the original word — that is, any single-character insertion, deletion, or substitution, or the transposition of two characters, that results in another real word. Thus none of the induced malapropisms were derived from closed-class words, and none were formed by the insertion or deletion of an apostrophe or by splitting a word. The data contained 1402 inserted malapropisms.

Because it had earlier been used for evaluating Mays *et al*'s trigram method, which operates at the sentence level, the dataset had been divided into three parts, without regard for article boundaries or text coherence: sentences into which no malapropism had been induced; the original versions of the sentences that received malapropisms; and the malapropized sentences. In addition, all instances of numbers of various kinds had been replaced by tags such as ⟨INTEGER⟩, ⟨DOLLAR_VALUE⟩, and ⟨PERCENTAGE_VALUE⟩. Actual (random) numbers or values were restored for these tags. Some spacing anomalies around punctuation marks were corrected.

2.2 Post hoc analysis of the data

A post hoc manual review of this dataset (after the evaluation described below was carried out) found that a number of the malapropized test items were unsuitable:

- **Errors in malapropism creation.** The malapropism-creation procedure included morphological analysis; the spelling variation operated on the base form and restored the morphology (so that *market's*, for example, could be replaced by *marker's*). This resulted in some overgeneration (e.g., *marketability* became *markerability*). Some other insertions were also non-words, apparently because of errors in the *ispell* lexicon (e.g., *advertiser* became *ladvertiser*; *suppliers* became *supliess*).

- **"Unfair" malapropisms.** Some of the malapropisms created were "unfair" in the sense that no automatic procedure could reasonably be expected to see the error. The canonical case is the substitution of *million* for *billion*, or vice versa; another is *employee* for *employer*, or vice

[1] *ispell* is a program that has evolved in PDP-10, Unix, and Usenet circles for more than 20 years, with contributions from many authors. Principal contributors to the current version include Pace Willisson and Geoff Kuenning.

versa, in many (but not all) contexts. In some cases, the substitution was merely a legitimate spelling variation of the same word (e.g., *labour* for *labor*). Additionally, some malapropisms were words so rare or obscure (e.g., *tunning* for *running*) that it seemed unreasonable to expect them to be present in Word's lexicon.

These cases were manually removed from the dataset. Determining what counted as "unfair" was obviously a judgment call. Words judged to be too "rare or obscure" were essentially those that were both absent from Word's lexicon and unknown to the author.

A total of 96 items were removed in this procedure, leaving 1306 "fair" items.

During this analysis, we observed that, because of the repetitive nature of the *WSJ* text, some malapropisms were repeated several times in the data (e.g., *chef executive officer*, *voice president*, *vice resident*). In addition, frequent words often recurred as the base word for different malapropisms (*money → monkey, honey, coney*; *share → sharp, shave, shame, shape, shark*).

3 Method

We presented the entire three-part dataset to Microsoft Office Word 2007 as a single text, and allowed it to perform a complete spelling check of the text. We selected the 'Ignore All' option for all non-word flags on words (mostly proper nouns) other than the induced malapropisms, except for those few instances that were genuine typos. We then recorded all flags for real-word errors, along with suggested corrections, and similarly for any non-word flags on the induced malapropisms.

4 Results

4.1 Quantitative data

We present results both for the complete, original dataset and for the dataset with "unfair" items removed (see section 2.2 above); we will refer to these as the *All Mals* and *Fair Mals* datasets respectively.

Table 1 shows recall, precision, and F for both datasets for *(a)* detection and *(b)* correction of the induced malapropisms. In the top part of the table, "Malapropisms detected as such", an induced malapropism was considered to be detected if it was flagged as a possible real-word error, and was considered to be corrected if detected and if the correct word appeared in the list of suggestions. In the lower part of the table, "Malapropisms detected as any error", an induced malapropism was considered to be detected if it was flagged as *either* a possible real-word error *or* a non-word error. That is, this part includes induced malapropisms that were not in Word's lexicon and were detected for that reason.

Recall was defined as the fraction of induced malapropisms detected or corrected. Precision was defined as the sum of the number of induced malapropisms detected or corrected (for each definition of detection) *plus* the number of false-positive real-word error flags from sentences without induced malapropisms, divided by the number of induced malapropisms. (See section 5.3 below for discussion of these definitions.)

Table 1: Performance measures for malapropism detection and correction by Microsoft Office Word 2007 on each dataset. TP = true positives, FP = false positives, R = recall, P = precision, $F = 2RP/(R+P)$.

Malapropisms detected as such

Data: All Mals (1402 items)

Detection: $TP = 310, FP = 11, R = .221, P = .966, F = .360$
Correction: $TP = 285, FP = 36, R = .203, P = .888, F = .330$

Data: Fair Mals (1306 items)

Detection: $TP = 308, FP = 11, R = .236, P = .966, F = .379$
Correction: $FP = 283, FP = 36, R = .217, P = .887, F = .349$

Malapropisms detected as any error

Data: All Mals (1402 items)

Detection: $TP = 347, FP = 11, R = .248, P = .969, F = .395$
Correction: $TP = 315, FP = 43, R = .225, P = .880, F = .358$

Data: Fair Mals (1306 items)

Detection: $TP = 334, FP = 11, R = .256, P = 968, F = .404$
Correction: $TP = 304, FP = 41, R = .232, P = .881, F = .367$

4.2 Observations

4.2.1 Analysis and examples

As the low recall shows, Word is very cautious in flagging real-word errors. Consequently, it missed more than three-quarters of the fair induced malapropisms. But when it did flag, it was almost always right in doing so. Moreover, except in cases subject to the bug discussed in section 4.2.3 below, its suggestions (usually there was only one) almost always included the correct word. Consequently, precision for both detection and correction are very high.

Table 2 shows some examples that Word detected and corrected, some examples in which Word detected the malapropism but failed to correct it, and some that Word failed to detect.

4.2.2 False positives and non-induced errors

In the entire 300,000-word text, Word flagged only 15 words, other than the induced malapropisms, as real-word errors. In three of these cases, it was correct (the error was in the original *WSJ* text); all were instances of spurious word separation (e.g., *through out* for *throughout*). In eleven cases, Word's flag was a false positive; a few examples are shown in table 3.[2]

4.2.3 A bug in Word regarding malapropisms in compound words

We observed that whenever a detected malapropism was the second half of a compound word (such as *last-minuet* and *price/earrings*), Word would almost always offer no suggestions for correction. (And the word would be displayed in red, not blue, in the 'Spelling and Grammar' window as if it were a non-word error, although the wavy underline in the main window would be blue.) This was so consistent (but not invariable; there was one exception in which suggestions were presented) that we assume it to be a bug in Word. Although it seems likely that, in the absence of the bug, Word could have offered correct suggestions in most of these cases, we were obliged to score them as instances of malapropisms detected but not corrected (there were 18 such cases). There was only one instance of a detected malapropism for which Word offered no suggestions that was not part of a compound word.

4.2.4 Non-word errors and the limitations of Word's lexicon

Some of the induced malapropisms were not in Word's dictionary and hence were detected as non-word errors; a few of these were such obscure words that we subsequently designated them as "unfair" (see section 2.2 above) (e.g., *tunning*). Nonetheless, Word has plenty of quite obscure words in its dictionary; and a few of the induced malapropisms that were flagged as non-words are surprising omissions (e.g., *cos, monte, coney*). There were 37 such cases (26 of which were in the "fair" data), and in 30 of them the correct word was in the suggestion list (21 of which were in the "fair" data). However, it was clear that Word's non-word corrector does *not* draw on the same mechanisms as the contextual corrector; there is no attempt to find the contextually best correction for non-words.

[2] In the fifteenth case (also a word separation case), it is impossible to determine whether Word's flag is correct; both the original and Word's suggested correction make sense in the context, albeit with different meanings. Ironically, the same sentence contained a genuine, but surely deliberate, malapropism that was not detected: in discussing a forthcoming U.S. presidential election, the author referred to then-president Ronald Reagan as the *recumbent*. Perhaps the two-letter substitution made it too distant for Word, as it would have for Wilcox-O'Hearn *et al* and Hirst and Budanitsky.

Table 2: Examples of Word's successes and failures in malapropism detection. The induced malapropism is shown in italics; the original word is in square brackets; Word's suggestion, if different, is shown in braces.

Malapropisms that Word successfully detected and corrected

Lyndon Johnson was crushed by Vietnam, practically driven from *officer* [office] and died a broken man.

In recent years, the intercity bus business has declined as air fares have plunged, and Greyhound has signaled the possible *salt* [sale] of its operations for some time.

Peter A. Cohen, chairman and chief executive officer, said, "The quarterly results highlight the diversity and strength of our *revenge* [revenue] stream."

A look back at Mr. Reagan's own public statements since the affair became public last fall indicates that many of the *resident's* [president's] comments have been uncertain, misleading or false.

Malapropisms that Word successfully detected but didn't correct

Also, they *rote* [note] {wrote} the company has acquired more than 23 million barrels of reserves, on an oil-equivalent basis, over the past three years.

Most consider it far more likely that MI23 had a *jot* [lot] {jolt} of right-wing nuts who saw reds under every bed.

William Jackman, a spokesman for the Air Transport Association, likened the new plan to "creating a three-*wane* [lane] {*no suggestions*} highway where there once was just one lane."

Malapropisms that Word didn't detect

Mr. Brady's aides apparently hope to *locus* [focus] attention solely on his public speech to the meeting, planned for tomorrow.

But the U.S. officials said the group in June was merely affirming the existing *rages* [ranges] and wasn't favoring an increase.

"We have been conducting a review of customer accounts and to the best of our knowledge customers have not lost *monkey* [money]," the spokesman said.

"The ministers and governors noted with satisfaction," the declaration said, "that the policies and commitments undertaken in the *curse* [course] of their cooperative efforts are producing results."

What the Europeans and House Democrats seem to *shark* [share] in common, however, is the apparent belief that if they close their eyes and wish hard enough, ...

Table 3: False positives flagged by Word. The flagged word, shown in italics, is correct; Word's suggestion is in braces.

But it's possible that what we're talking about was use of money to pay people and hire individuals who could *effect* {affect} a rescue of our people there.

All of this comes as news to me — and probably *to* {too} many Americans.

The all-cash planned acquisition also will *heap* {hear} a lot of debt on Southdown, which just last year had to work out from under a heavy debt load after buying back about 56% of its shares.

Table 4: A comparison of recall, precision, and F for three methods of malapropism detection on the All Mals dataset.

Word (real-word errors)	
Detection: $R = .221, P = .966, F = .360$	Correction: $R = .203, P = .888, F = .330$
Word (all errors)	
Detection: $R = .248, P = .969, F = .395$	Correction: $R = .225, P = .880, F = .358$
Trigrams	
Detection: $R = .544, P = .528, F = .536$	Correction: $R = .491, P = .503, F = .497$
Lexical cohesion	
Detection: $R = .306, P = .225, F = .260$	Correction: $R = .281, P = .207, F = .238$

5 Discussion

5.1 Comparison with other methods

Table 4 shows Word's results on the All Mals dataset (repeated from table 1) compared with the results for the trigram method (by Wilcox-O'Hearn *et al* 2008)[3] and the lexical cohesion method (by Hirst and Budanitsky 2005). (Results are not available for these methods on the Fair Mals dataset.)

As can be seen, the trigram method performs notably better than Word, which in turn performs notably better than lexical cohesion.

5.2 A fair test?

This experiment would be unfair to Word if Word uses any context beyond the sentence (as Hirst and Budanitsky (2005) do), because the malapropism section of the test text was just a big pile of sentences with malapropisms in each one, without the original context of the sentence. Ideally, some or all of the evaluation should be repeated with the malapropisms in place in the original articles in order to see if this improves Word's performance, but the data would need to be recreated in order to do this.

The experiment is also unfair in not testing all of Word's abilities; possibly Word could have improved its score if the test data had covered additional kinds of real-word errors. The absence of split-word errors and apostrophe errors from the test data has already been noted. Also untested were errors with an edit distance greater than 1 that are nonetheless likely in practical use. These include phonetic confoundings such as *cymbal* / *symbol* and *spayed* / *spadé*[4] and standard word confoundings such as *there* / *their* and *principle* / *principal*.[5]

[3] The data shown here are not from this paper, but rather are later results following the correction of a minor bug and with a slight change to the method that removed the distinction, made previously, between upper- and lower-case characters.

[4] Toutanova and Moore (2002) present a model of spelling variation based on phonetic confounding.

[5] *There* / *their* is explicitly mentioned in Microsoft's online guide to Word as an example that Word can correct. `http://office.microsoft.com/en-us/word/HA100742241033.aspx`, `http://office.microsoft.com/en-us/word/HP101194671033.aspx`.

5.3 Defining recall and precision for this task

It was not straightforward to define precision and recall for this task in a way that provides a reasonable measurement of real-word error detection and correction.

The wholly conventional method is to define recall as the fraction of induced malapropisms flagged, or flagged and corrected, and precision as the fraction of flagged words, or flagged and corrected words, that are induced malapropisms. One problem with this, which is easily dealt with, is that a few of the flagged and corrected words that were not induced malapropisms were real-word errors nonetheless. Clearly, Word should not be penalized for finding these, so they are discarded from the data. The data using these definitions of recall and precision are shown in the top half of table 1.

A more-difficult problem is the extent to which Word should be given credit for finding induced malapropisms that it thinks are non-words. From a functional perspective, Word has succeeded in such cases by finding an error and bringing it to the user's attention, even if it didn't do so "the right way", and hence it deserves some credit. But the problem then arises as to how to treat all the other words flagged as non-word errors, most of which are not true errors. It seems unfair to penalize Word's precision score for false positives here (especially for those that are proper nouns); they can hardly be considered to be false positives in malapropism detection. And yet they arise from the same cause, gaps in Word's lexicon, that leads to Word's "successes" in flagging induced malapropisms as non-words, so to ignore them results in a situation in which omissions from Word's lexicon are possibly rewarded and never penalized. Nonetheless, the lower half of table 1 shows data computed with these definitions; induced malapropisms flagged as non-words are counted as true positives, but nothing else flagged as a non-word is counted as a false positive. This data should be interpreted cautiously, as it rewards lexical inadequacy; in the limit, a system that simply flagged every word as a non-word error would score $R = P = F = 1.0$!

6 Conclusion

The contextual spelling corrector in Microsoft Office Word 2007 is a cautious (low recall) but believable (high precision) system. However, its overall performance, as measured by F, is much poorer than that of the trigram method of Mays *et al* (1991).

The trade-off between the two systems is a difficult one. In simple terms, better performance is better; but believability is an important attribute for a consumer-level system ("if Word says it's wrong then it's wrong") and could well be considered worth sacrificing performance for.[6] The problem with this, however, is that as users become familiar with the system, their expectations will rise and believability will start to apply also to what Word fails to flag ("If Word says it's right then it's right"). A system that is more visibly error-prone might actually serve users better.

[6]This is a greater issue for malapropism detection than non-word detection, as it is easier for the typical to user to understand the limitations of the simple list-based non-word method; each time a correctly spelled name is flagged, for example, the user sees that "it's okay, it's just that Word doesn't know that word".

References

Hirst, Graeme and Budanitsky, Alexander (2005). Correcting real-word spelling errors by restoring lexical cohesion. *Natural Language Engineering*, 11(1), 87–111.

Hirst, Graeme and St-Onge, David (1998). Lexical chains as representations of context for the detection and correction of malapropisms. In: Fellbaum, Christiane, (editor), *WordNet: An electronic lexical database*, Cambridge, MA:, The MIT Press, 305–332.

Mays, Eric; Damerau, Fred J.; and Mercer, Robert L. (1991). Context based spelling correction. *Information Processing and Management*, 23(5), 517–522.

Microsoft Corporation (2006). *Microsoft Office Word 2007* [product guide]. Downloadable from `http://office.microsoft.com/en-us/word/HA101680221033.aspx`

Toutanova, Kristina and Moore, Robert C. (2002). Pronunciation modeling for improved spelling correction. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, 144–151.

Wilcox-O'Hearn, L. Amber; Hirst, Graeme; and Budanitsky, Alexander (2008). Real-word spelling correction with trigrams: A reconsideration of the Mays, Damerau, and Mercer model. *Proceedings, 9th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2008)*, Haifa, 605–616.