Real-word spelling correction with trigrams:
A reconsideration of the Mays, Damerau, and Mercer model

Amber Wilcox-O'Hearn, Graeme Hirst, and Alexander Budanitsky
University of Toronto, Department of Computer Science
amber, gh, abm @cs.toronto.edu

Abstract
The trigram-based noisy-channel model of real-word spelling-error correction that was presented by Mays, Damerau, and Mercer in 1991 has never been adequately evaluated or compared with other methods. We present a new evaluation that enabled a meaningful comparison with the WordNet-based method of Hirst and Budanitsky (2005) and the "contextual spelling corrector" of Microsoft Office Word 2007. The trigram method was found to be superior to both these other methods, even on content words. We also found that optimizing over sentences gives better results than variants of the algorithm that optimize over fixed-length windows.

1. Real-word spelling errors

Real-word spelling errors (malapropisms)

- Can't be detected by regular (lexicon-based) spelling checkers.

Methods

1. Predefined *confusion sets* of common errors (Golding & Roth 1999) — e.g., *principal / principle*:
   - Choose most-likely member in context.
   - *Limitation:* Can only deal with predefined errors.
2. *Cohesion-based* (Hirst & Budanitsky 2005):
   - Use WordNet to find relationships in text.
   - Words unrelated to context are *semantic anomalies;* replace with spelling variation that *is* related — e.g., … *months in the ~~pear~~ year* … because *month / year* are related, *month / pear* are not.
   - *Limitation:* Works only on content words.
3. *Trigrams* (Mays, Damerau, & Mercer 1991) [**MDM**]:
   - Try to increase trigram probability of sentence by replacing words with spelling variations [see box 2].

Which method is best?

- **Trigram method has never been evaluated in comparable terms.**
- We replicated it to evaluate it and try to improve it.

2. MDM's trigram method

Noisy channel model

- Typist model:
  - Errors introduced by typist at rate $(1-\alpha)$; distributed equally among set $SV(w)$ of *spelling variations* of each word w.
  - $SV(w)$ defined by edit distance and/or phonetic similarity.
  - Probability that word w is typed as x:

$$P(x|w) = \begin{cases} \alpha & \text{if } x = w \\ (1-\alpha)/|SV(w)| & \text{if } x \in SV(w) \\ 0 & \text{otherwise} \end{cases}$$

- Language model uses *trigram probabilities*:
  - Probability of intended sentence:

$$P(S') = \prod_{i=1}^{n+2} P(w_i | w_{i-1} w_{i-2}).$$

Sentence correction

- Generate candidate correction sentences: substitute word with spelling variation.
- Corrected sentence maximizes

$$P(S'|S) \propto P(S') \cdot P(S|S')$$

where $P(S|S')$ is the typist model.

Advantages

- Detects errors in both content and function words.
- Spelling variations need not be predefined.

Disadvantages

- Language model is large.
- Data sparseness is a problem.
- Corrects at most one word per sentence.

3. Why re-evaluate MDM's method?

MDM's results can't be compared to those from other methods

- Used *unnamed corpus;* vocabulary only 20,000 words.
- Highly *imbalanced test set*: Only 100 correct sentences but 8600 erroneous sentences; no OOVs.
  - Potential false positives extremely small compared to potential true positives.
  - Hence claim of per-sentence precision > .99 and recall of .618 to .744 is meaningless.

Results presented in *incommensurate* terms

- Only per-sentence detection and correction for erroneous sentences; false positives for correct sentences.
- Conflated correction of wrong word with wrong correction of right word.
- ***No*** per-word accuracy, precision, recall, or false/true positives.

So replicate, and gather data in commensurate terms.

Enable comparison with cohesion-based method (Hirst & Budanitsky) and with MS Office Word 2007.

4. Our re-evaluation

- *Training data:* 1987–89 *WSJ* corpus (30M words).**
- *Language model:* (a) 20,000 word vocabulary*, and (b) 62,000-word vocabulary; other words mapped to OOV token; Kneser-Ney smoothing.
- *Test data:* 500 reserved *WSJ* articles (300,000 words, 15,555 sentences).
  - Replaced 1 word in every 200 with real-word error (*i.e.*, α = .995).**
  - Did this 3 ways (created 3 test sets):
    - **T20:** Any word replaced with a spelling variation from 20K vocabulary model.*
    - **T62:** Any word replaced with a spelling variation from 62K vocabulary model.
    - **MAL:** Any content word from WordNet replaced with a spelling variation from *ispell*.**

* Replicates MDM's evaluation.
** Replicates Hirst & Budanitsky's evaluation.

Results improved over those in proceedings paper!
Due to better language model, better handling of case, bug fixes, etc.

Most realistic condition: Typist is 99.5% accurate.

5. Results (with 62,000-word vocabulary)

α	Detection P	Detection R	Detection F	Correction P	Correction R	Correction F
Test data T20:						
.9	.331	.853	.477	.324	.829	.466
.99	.562	.775	.651	.556	.756	.641
.995	.635	.738	.683	.629	.722	.672
.999	.771	.656	.709	.768	.643	.700
Test data T62:						
.9	.340	.882	.491	.333	.851	.478
.99	.581	.828	.683	.573	.804	.670
.995	.656	.795	.719	.650	.775	.707
.999	.796	.740	.767	.792	.724	.757
Test data MAL:						
.9	.252	.664	.365	.243	.633	.351
.99	.457	.583	.513	.448	.563	.499
.995	.531	.550	.540	.524	.534	.529
.999	.692	.484	.569	.687	.472	.560
Lexical cohesion method (on MAL):						
	.225	.306	.260	.207	.281	.238
Microsoft Office Word 2007 (on MAL):						
Strict scoring						
	.966	.221	.360	.888	.203	.330
Generous scoring						
	.969	.248	.395	.880	.225	.358

Trigram method beats lexical cohesion and Word both for detection and for correction.

6. Try to improve the method

Want possibility of more than one correction in a single sentence.

- Not possible in original method: combinatorially explosive.

New methods

Instead of using the single best overall correction:

1. Combine ***all*** corrections that improve the overall sentence probability. (Might not improve overall probability when combined.)

   or

2. Combine single best correction from ***smaller fixed-length windows***.

Results

- In all cases, performance never improved, and was often worse.
- ***Reason:*** Reduced precision because of marked increase in false positives.

Conclusion: Limit of one correction per sentence is a useful constraint.

7. Conclusion

Related research

- Noisy channel trigram models are also used in the simpler problem of nonword spelling correction, with an emphasis on improved channel models; e.g.
  - character-based confusion sets to model typing errors as they occur in practice (Church and Gale 1991).
  - edit distances based on phonetic similarity (Toutanova and Moore 2002).

Our next step

- Extend the present MDM model to use Church and Gale's (1991) model of typing errors (Wilcox-O'Hearn 2008).

Bottom line

The noisy-channel trigram method of real-word spelling correction is superior to both the cohesion-based method and the proprietary method of Microsoft Office Word 2007.

References

Church, Kenneth W. and William A. Gale (1991). Probability scoring for spelling correction. *Statistics and Computing*, 1, 93–103.

Golding, Andrew R. and Dan Roth (1999). A Winnow-based approach to context-sensitive spelling correction. *Machine Learning*, 34(1–3), 107–13.

Hirst, Graeme (2008). An evaluation of the contextual spelling checker of Microsoft Office Word 2007. *http://www.cs.toronto.edu/compling/Publications/publications.html*

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

Kukich, Karen (1992). Techniques for automatically correcting words in text. *Computing Surveys*, 24(4), 377–439.

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

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

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

Wilcox-O'Hearn, L. Amber (2008). *Applying trigram models to real-word spelling correction*. MSc thesis, Department of Computer Science, University of Toronto [forthcoming].

UNIVERSITY of TORONTO

NSERC CRSNG Financially supported by the Natural Sciences and Engineering Research Council of Canada

Presented at the 9th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2008), 17–23 February 2008, Haifa, Israel.