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

Amber Wilcox-O'Hearn, Graeme Hirst, and Alexander Budanitsky *

Department of Computer Science, University of Toronto
Toronto, Ontario, Canada M5S 3G4
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 analyze the advantages and limitations of the method, and present a new evaluation that enables a meaningful comparison with the WordNet-based method of Hirst and Budanitsky. The trigram method is found to be superior, even on content words. We then show that optimizing over sentences gives better results than variants of the algorithm that optimize over fixed-length windows.

1 Introduction

Real-word spelling errors are words in a text that, although correctly spelled words in the dictionary, are not the words that the writer intended. Such errors may be caused by typing mistakes or by the writer's ignorance of the correct spelling of the intended word. Ironically, such errors are also caused by spelling checkers in the correction of non-word spelling errors: the "auto-correct" feature in popular word-processing software will sometimes silently change a non-word to the wrong real word (Hirst and Budanitsky 2005), and sometimes when correcting a flagged error, the user will inadvertently make the wrong selection from the alternatives offered. The problem that we address in this paper is the automatic detection and correction of real-word errors.

Methods developed in previous research on this topic fall into two basic categories: those based on human-made lexical or other resources and those based on machine-learning or statistical methods. An example of a resource-based method is that of Hirst and Budanitsky (2005), who use semantic distance measures in WordNet to detect words that are potentially anomalous in context — that is, semantically distant from nearby words; if a variation in spelling [1] results in a word that was semantically closer to the context, it is hypothesized that the original word is an error (a "*malapropism*")

* This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. We are grateful to Bryce Wilcox-O'Hearn for his assistance.

[1] In this method, as in the trigram method that we discuss later, any consistent definition, narrow or broad, of what counts as the spelling variations of a word may be used. Typically it would be based on edit distance, and might also take phonetic similarity into account; see our remarks on Brill and Moore (2000) and Toutanova and Moore (2002) in section 5 below.