Reviewer's report

Title: Doublet method for very fast autocoding

Version: 1 Date: 15 August 2004

Reviewer: David Johnson

Reviewer's report:

In general a well written easy to follow paper. The major weakness of the paper is that the citations should be expanded. There has been much related work in the fields of information retrieval and natural language processing that deal with the problems of text matching. Furthermore, the very important issues and questions, raised by the author, surrounding the evaluation of information systems and the use of precision and recall as a means of evaluation, have been better addressed in the Information Sciences – while Tim Bray's online article touches upon the issues there are much better scholarly sources (e.g., Journal of the American Society for Information Science, ACM SIGIR). Ultimately, the utility (e.g., accuracy, "relevance") of a medical autocoder must be examined.

While the execution time comparison between the doublet method and the phrase match provides interesting anecdotal evidence, it should be noted that the speed differences may have more to do with implementation details rather than the actual complexity of the two algorithms. While the fact that the two algorithms were implemented using Perl does eliminate one variable, other implementation factors (e.g., data structures used) can significantly impact speed.

The code snippets are helpful to understand the algorithm, although not every reader may be familiar with Perl. It may improve the readability if the author simplifies the code examples using a form of pseudo code.

Finally, the paper would be well served if the author discussed cases where the algorithm fails. For example, how does the method handle sentence boundaries? Or if two overlapping candidate phrases in a text have one or more common doublets – which phrase wins?

What next?: Accept after discretionary revisions

Level of interest: An article whose findings are important to those with closely related research interests

Quality of written English: Acceptable

Statistical review: No

Declaration of competing interests:

None