# Errors in wikis: new challenges and new opportunities — a discussion document

**Ann Copestake**
Computer Laboratory
University of Cambridge
`aac@cl.cam.ac.uk`

## Abstract

This discussion document concerns the challenges to assessments of reliability posed by wikis and the potential for language processing techniques for aiding readers to decide whether to trust particular text.

## 1 Wikis and the trust problem

Wikis, especially open wikis, pose new challenges for readers in deciding whether information is trustworthy. An article in a wikipedia may be generally well-written and appear authoritative, so that the reader is inclined to trust it, but have some additions by other authors which are incorrect. Corrections may eventually get made, but there will be a time lag. In particular, many people are now using Wikipedia (`www.wikipedia.org`) as a major reference source, so the potential for misinformation to be spread is increasing. It has already become apparent that articles about politicians are being edited by their staff to make them more favourable and no doubt various interest groups are manipulating information in more subtle ways. In fact, as wikis develop, problems with reliability may get worse: authors who wrote an article several years ago won't care so much about its content and may not bother to check edits. When obscure topics are covered by a wiki, the community which is capable of checking facts may be small.

Of course errors arise in old text too, but a generally authoritative conventional article is unlikely to contain a really major error about a central topic. Different old text publications have different perspectives, political or otherwise, but the overall slant is usually generally known and hence not problematic. Non-wiki web pages may have unknown authors, but the domain offers some guide to reliability and to likely skew and the pages can be assessed as a whole. The issue here is not the overall number of errors in wikis versus published text or web pages, but how a reader can decide to trust a particular piece of information when they cannot use the article as a whole as a guide.

There is a need for automatic tools which could provide an aid for the reader who needs to assess trustworthiness and also for authors and moderators scanning changes. Similarly, moderators need tools for identification of vandalism, libel, advertising and so on.

Questions:

1. Is wiki reliability really a problem for readers, as I hypothesise? Perhaps readers who are not expert in a topic can detect problematic material in a wiki article, despite the multiple authorship.

2. Can we use language processing tools to help readers identify errors and misinformation in wiki pages?

## 2 Learning trustworthiness

The availability of change histories on wikis is a resource which could be exploited for training purposes by language processing systems designed to evaluate trustworthiness. If it is possible to categorise users as trustworthy or non-trustworthy/unknown by independent criteria (such as overall contribution level), then we can use changes made by trustworthy users that delete additions made by the unknown users as a means of categorising some text as bad. (Possibly the

comments made by the editors could lead to sub-categorization of the badness as error vs vandalism etc.) A tool for highlighting possible problem edits in wikis might thus be developed on the basis of a large amount of training data. Techniques derived from areas such as language-based spam detection, subjectivity measurement and so on could be relevant. However, one of the relatively novel aspects of the wiki problem is that we are looking at categorisation of small text snippets rather than larger quantities of text. Thus techniques that rely on stylistic cues probably won't work. Ideally, we need to be able to identify the actual information provided by individual contributors and classify this as reliable or unreliable. One way of looking at this is by dividing text into factoids (in the summarisation sense). Factoid identification is a really hard problem, but maybe the wiki edits themselves could help here.

Questions:

1. Can we automatically classify wiki contributors as reliable/unreliable?

2. Do trustworthy users' edits provide good training data?

3. Are there any features of text snippets that allow classification of reliability? (My guess: identification of vandalism will be possible but more subtle effects won't be detectable.)

4. What tools could be adapted from other areas of language processing to address these issues?

## 3 An ontology of errors?

As an extension of the ideas in the previous section, perhaps wiki histories could be mined as a repository of commonly believed false information. For instance, the EN wikipedia entry for University of Cambridge currently (Jan 5th, 2006) states:

> Undergraduate admission to Cambridge colleges used to depend on knowledge of Latin and Ancient Greek, subjects taught principally in the United Kingdom at fee-paying schools, called public schools.
> ('public schools' was linked)

One way in which this is wrong is that British 'public schools' (in this sense) are only a small proportion of the fee-paying schools, but equating public schools with all fee-paying schools is a common error. Suppose a trustworthy editor corrects this particular error in this article (and perhaps similar errors in the same or other articles). If we can automatically analyse and store the correction, we could use it to check for the same error in other text. As wikis get larger, this might become a useful resource for error detection/evaluation of many text types. Thus errors in wikis are an opportunity as well as a challenge.