

Author's response to reviews

Title: Preparation of name and address data for record linkage using hidden Markov models

Authors:

Dr Tim Churches (tchur@doh.health.nsw.gov.au)

Dr Peter Christen (peter.christen@anu.edu.au)

Kim Lim (klim@doh.health.nsw.gov.au)

Justin X Zhu (u3167614@student.anu.edu.au)

Version: 2 Date: 8 Dec 2002

The authors thank the reviewers for their detailed, thoughtful and very helpful comments.

Discretionary revisions requested by Reviewer 1:

1. Better discussion of emission probabilities:

Additional explanation and definitions of the transition and emission matrices have been added to the section titled "Hidden Markov models" starting on page 10.

2. Illustration of the models:

Tables 7 and 8 and Figures 2 and 3 have been added to give the reader an idea of the models currently implemented in the Febrl package. Figure 3 is only really useful to convey an impression of the complexity of typical HMMs!

3. Run-time performance issues:

New material has been added to the Methods, Results and Discussion sections on the run time performance of the HMMs as implemented in Python, on pages 20, 21 and 22 respectively. At this stage of the development of the Febrl package, we have resisted the temptation to engage in "premature optimisation".

4. "Bootstrap" training:

We have removed references to the term "bootstrap" to avoid confusion and replaced it with the term "iterative refinement" as suggested by Reviewer 2. "Bootstrap" was being used in the computer science sense, rather than the statistical sense. The more training records used, the better, but we found that as few as 1500 was sufficient. The iterative refinement training method merely reduces the mental effort and the number of keystrokes which the trainer is required to make, but every record in the training set still needs to be checked by the trainer. However, we found that after the first one or two iterations, the majority of records were tagged correctly and the training task became one of scanning the tagged records for exceptions, and correcting these. The use of null models to detect "poorly fitting" records, mentioned at the end of the Discussion section, promises to make this task more efficient, but we have not had time to incorporate this into the Febrl package as yet. We found that the HMMs were indeed quite robust despite numerous, and substantial expansions of the look-up tables while we were developing the software.

5. Comparison of run-time performance with AutoStan: Directly comparable run-time performance for AutoStan on the same task on the same file and the same computer has been added to the Results section.

Discretionary revisions requested by Reviewer 2

Tokenisation: The description of the approach used by Borkar et al. has been corrected. We also more fully acknowledge the influence which the work of Borkar et al. had on the development of the HMM pre-processing aspects of the Febrl package, as described in the paper.

Modeling: The models used were developed heuristically, based on previous experience with the AutoStan package - the manuscript has been modified to make this clear (last paragraph, page 14). As noted above, tables and figures have been added which define and visualise the actual models used. Also noted in the manuscript, future versions of the Febrl package will feature "soft-coded" models which will allow end users to define their own models. The automatic model selection procedure described by Borkar et al. is elegant and attractive, but we feel that because the models ultimately need to be trained by humans, they must make intuitive sense, and thus heuristic methods based on empirical knowledge of local name and address forms are a satisfactory (and much simpler) approach.

Training: We have adopted Reviewer 2's terminology of "iterative refinement" in order to avoid confusion over our use of the term "bootstrap", as noted above. The training of the name standardisation model, described on pages 18 and 19, was similar to that used for the address model. However, the evaluation of the performance of the name standardisation included a 10-fold cross-validation. Such a cross-validation is de rigueur when reporting on machine-learning classification techniques, and is used to ensure that "over-fitting" has not been used to artificially improve classification accuracy (the bias-variance trade-off). However, as noted in the manuscript, "over-fitting" of HMMs to specific data sets to be processed is likely to be the norm in most research projects. In retrospect, better co-ordination between the authors on the way in which the performance of the name and address standardisation was reported would have been desirable. Unfortunately, available resources do not permit the evaluation of the name standardisation to be redone at this stage.

At this stage, we are unable to devote more time to refining the name standardisation model - we need to concentrate on other aspects of the overall Febrl record linkage package - but the discussion provided by Reviewer 2 on this topic will be extremely useful in informing improvements when we (or someone else, since the code is open source) have time to revisit the name standardisation problem, and we are most grateful for this generous contribution.

Tim Churches, Peter Christen, Kim Lim and Justin Zhu.