Reviewer's report

Title: A Controlled Trial of Automated Classification of Negation from Clinical Notes

Version: Date: 6 October 2004

Reviewer: Jeremy E Rogers

Reviewer's report:

General

None.

--------------------------------------------------------------------------------

Major Compulsory Revisions (that the author must respond to before a decision on publication can be reached)

The initial introduction of this paper is not coherent. In addition to some minor typos, it includes individual sentences or larger paragraph sections that appear more than once.

The first para of the introduction notes that several groups have applied NLP techniques, however the phrasing might be taken as suggesting that they have been largely successful in this pursuit when, as far as I know, NLP remains as problematic as the previous sentence acknowledges. Half the references cited as evidence that NLP has been used by many groups (nos 4-10) refer to the work of a group of researchers known mainly for their work in concept-oriented clinical terminology design and not NLP (eg Price, Schulz and Chute). These citations should be replaced or augmented with references to the body of NLP and information extraction literature, especially those outside the medical informatics community.

The section following, presenting the history of classification and of a compositional approach more specifically, seems overlong and irrelevant to the experiment being reported. Similarly the penultimate two paragraphs of the conclusion do not seem relevant, unless the authors’ point is that information extraction from text remains a very attractive route to obtaining large volumes of good quality structured clinical information, but that cracking the nut of negation is one important barrier to achieving this. If that is the point, it should be made more clearly and economically.

The presentation of the results is unclear: how were the 14,792 concepts identified? Is this 14,792 discrete concepts, or 14,792 concept occurrences (but where one concept may appear, and be counted, more than once)? 2028 negative concepts are reported as having been found, but 205 did not have a SNOMED mapping and, therefore, were presumably not found by automatic means. How were they found?

Whilst the measured performance of this intervention, averaged across the experimental corpus, does appear to be as good as or better than other reported similar interventions, there is as yet little research evidence that provides us with a figure for what levels of accuracy in information extraction from clinical text are 'good enough'. I am unaware of any research defining adequate input data accuracy for decision support software, which the opening paragraph acknowledges as a significant motivator for the work presented. But, in the absence of such metrics, the performance of this information extraction intervention can not be judged by the same criteria as a diagnostic laboratory test, where the consequences of false results are entirely different. Any explicit or implied extrapolation of fitness for purpose(s) from the observed performance figures needs careful
justification, and where no prior data exists to indicate what accuracy is required for a given application, the authors should be cautious in claiming any fitness for that purpose.

The interpretation of results is interwoven with the presentation of those results and the conclusion, and would be better in a separate paragraph or section. The phraseology of interpretation (e.g. 'quite excellent', 'reliably') should be improved, and all interpretations justified or reconsidered in the light of the comments above.

The headline average figures of 97.2% sensitivity and 98.8% specificity quoted in the abstract should be qualified to reflect the range of values returned in the subanalysis of performance by section of the medical record, as reported in Table 2.

-------------------------------------------------------------------------------

Minor Essential Revisions (such as missing labels on figures, or the wrong use of a term, which the author can be trusted to correct)

In the methods section it is not clear to me whether the experimental sample was 41 discrete documents from 41 patients, or the entire patient record for each patient, where each record would be comprised of more than one document. The abstract suggests it was one document per patient.

-------------------------------------------------------------------------------

Discretionary Revisions (which the author can choose to ignore)

The methods and conclusions sections mention an 'ontology of terms that start negation' but it isn't clear in what respect this is an ontology, and how it differs from a lexicon or catalogue of the terms. Perhaps this is reported in another paper. If not, some detail could usefully be provided to compensate for the necessary trimming of the introduction.

The conclusions section reports that one common reason for failure was that SNOMED could not represent the negative concepts. If I understand the implications of this statement correctly, the authors could acknowledge the current debate as to whether SNOMED should contain ANY capability to represent negated forms of concepts, with negation instead always handled exclusively in the information model and not, as now, possibly in both.

Although I may accept the argument that identifying negation of concepts in text is trivial and likely to be repeatable, and therefore that the results of a single reviewer are likely to be reasonably reliable, I am less confident that the categorisation of the reason for any failure is as reproducible. Such a categorisation may be especially problematic given that the distinction between a Qualifier and a Modifier is contentious in the clinical terminology community. However, as the paper does not make much of the failure analysis, this is less of a problem.

What next?: Unable to decide on acceptance or rejection until the authors have responded to the major compulsory revisions

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

Quality of written English: Needs some language corrections before being published

Statistical review: No

Declaration of competing interests:

None