Reviewer's report

**Title:** Recognition of medication related entities from discharge summaries using ensembles of classifiers

**Version:** 6 **Date:** 2 December 2011

**Reviewer:** Ira Goldstein

**Reviewer's report:**

The authors present an ensemble classification approach to a named entity recognition task using medical records. They employ two complementary classification approaches: support vector machines and conditional random fields. They claim that an ensemble of classifiers can perform better than individual classifiers on a named entity recognition task.

**Major Revisions**

While fairly well written, the paper could be improved by a professional editor.

1. As written, the authors still present unsubstantiated claims. Under Results and discussions, the authors use the terms such as “substantially”, “comparable”, and “better” to describe differences in performance. Differences in performance must be noted as being either statistically significant or not statistically significant at a given level of confidence. Without this information, we do not know if the differences noted are true differences or if they are due to random chance.

Since Table 9 comes after Tables 8 & 9, it would appear as if statistical significance is not provided for the 10-Fold results. For example, on page 15 the authors state:

“Another interesting finding from Table 5 is that while the SVM-based system is better than the CRF-based system in medication.” I would question if the difference between 92.13 and 92.26 is significant.

Additionally, even when the authors present statistical significance in their tables, the findings appear to be ignored in the narrative. For example, even though table 9 shows that the differences between the Sydney system and each of the CRF and SVM systems are “NS”, on page 16 the authors state:

“Table 7 shows that compared to the Sydney team’s result (F-score of 89.19%), our methods are better, but comparable, with F-score of 89.26% from the CRF, and 89.21% from the SVM. This shows that using customized MedEx as inputs to the SVM and CRF helps to improve the F-score of those two machine learning methods.”

If the differences are not significant then none of the systems can claim to be better than any of the others. Even if significant, since the Sydney system used CRF, it is not clear how this would show that customized MedEx as an input
would improve the performance of an SVM system.

Statistical significance needs to be provided for each of the tables presenting performance results. Rather than creating a separate table, statistical significance can be shown in the body of the table by using either asterisks or bold text.

2. The reason provided for employing simple majority voting is “that it works well when individual classifiers have different characteristics with regard to the precision and recall.” It is not apparent from the tables that this is the case.

3. The first sentence in the explanation for Local SVM-based voting appears to contradict the second sentence.

- Minor Revisions

1. The authors introduce the six i2b2 entities on page 4 as “medication information” and “fields.” However, subsequent mentions use the terms “entities” and “classes” without bridging the several terms. I would pick one way to collectively refer to the six and use it consistently in the paper.

2. Page 10 states: “Words only. We refer to it as the baseline method in this study.” Yet on page 15 the baseline refers to the customized MedEx system.

3. On page 5 the author state that:
“It is likely that the richer training data used by the Sydney system was responsible for the difference in performance between those two systems,” but neither provide a reason why this is likely nor explore any differences between the two systems.

4. On page 6, the text “Clearly the Sydney team developed their approach during the i2b2 challenge period under time pressure while that was not an issue for us” does not flow from the prior sentence.

5. On page 7, the term “BIO format” is presented without explanation.

6. Citations 21 and 22 have misspelled Özlem Uzuner’s name. The umlaut is in her first name, not her last name.

- Discretionary 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:** Yes, and I have assessed the statistics in my report.
Declaration of competing interests:

I have no competing financial interests.

I work with the organizers of the i2b2 shared task who provided the data for this paper.