On semi-automated matching and integration of database schemas

Ünal Karakaş, Ö.

Publication date
2010

Appendix E

Evaluation of Schema Matching – For "select max above threshold" strategy

In the second type of experiments, we used the "select max above threshold" strategy. Results of precision, recall, f-measure, and overall measures for SASMINT and COMA++ are shown in Figures E.1 through E.8. In general, this strategy achieves better than the "select all above threshold" strategy, when precision, f-measure, and overall are considered. This is due to the fact that not all matches above the threshold are selected, but only those with higher similarity values. This brings about less number of false positives, which means precision is higher than the "select all above threshold" strategy. However in some cases, this strategy may lead to lower recall values because of missing some correct matches. This effect is little compared to the high increase in precision. Therefore, in general, the values for f-measure and overall were higher for both SASMINT and COMA++, when "select max above threshold" strategy was used, but again on average SASMINT performed slightly better then COMA++ .

E.1 Evaluation of Schema Matching Using Precision

Precision values for SASMINT and COMA++ were the same for the purchase order, UNIV-2, and UNIV-3 schemas. For the purchase order and UNIV-3 schemas, they both had the precision of 1.0. For hotel and UNIV-1 schemas, SASMINT performed around 1.08 times better than COMA++, but for the SDB schema, COMA++ had 1.05 times higher precision value. The reason for the difference between the performances of the two systems was because of the false positives introduced by the systems. For example, for the SDB schema, SASMINT identified "donorID" and "donorVisitID" as similar, which was incorrect and thus resulted in a decrease in the precision. On the average, SASMINT achieved 0.94 precision over all schema pairs, whereas the average precision of COMA++ was 0.93. Compared to the "select all above threshold" strategy, precision of "select max above threshold" strategy was very high for both systems. This is due to the fact that, in this second strategy only the most relevant matches were selected, which had the higher similarities than irrelevant matches. For example in the test with UNIV-2, although the first strategy identified "STAFF_EMAIL" column of the "academic_staff_member" table and "ELECTRONIC_MAIL" column of the "university_student" table as similar, the second strategy did not make this mistake. Figures E.1 and E.2 show the complete results for COMA++ and SASMINT.

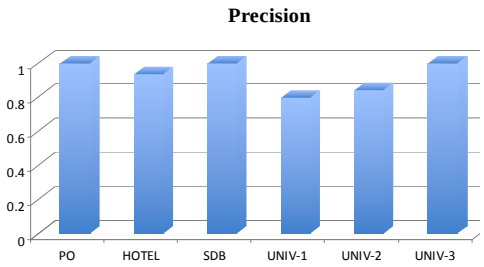

Fig. E.1. Precision values for **COMA++** - select max above threshold strategy

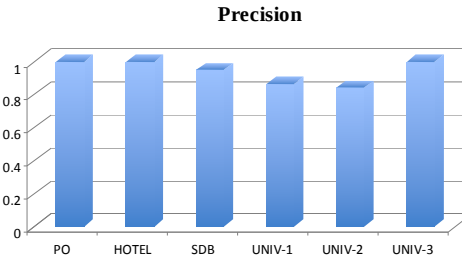

Fig. E.2. Precision values for **SASMINT** - select max above threshold strategy

E.2 Evaluation of Schema Matching Using Recall

Recall values in the case of "select max above threshold" strategy were either the same or a bit lower than the ones in the "select all above threshold" strategy for SASMINT. For COMA++, recall was lower in the "select max above threshold" strategy for all schema pairs. This was because of missing some correct matches. SASMINT and COMA++ both had the same recall values for the SDB and UNIV-2 schemas, as shown in Figures E.3. and E.4. For the hotel schemas, COMA++ was 1.1 times better than SASMINT. This was because of some table-to-table and column-to-column matches that could not be identified by SASMINT. Namely, similar to the case in "select all above threshold" strategy, these matching pairs were semantically similar, but since the current version of WordNet did not provide high "semantic" similarity values for these pairs, SASMINT could not identify them as similar. However, SASMINT achieved for the purchase order and UNIV-3 schemas 1.2 times and for the UNIV-1 schemas 1.1 times better than COMA++. On average, SASMINT had recall of 0.77, whereas COMA++ had 0.72.

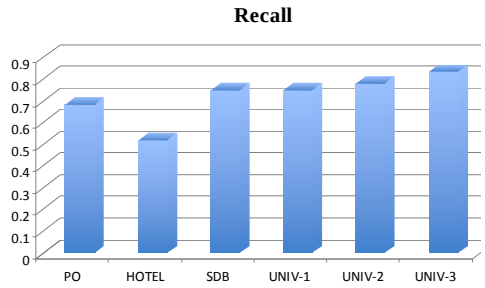

Fig. E.3. Recall values for **COMA++** - select max above threshold strategy

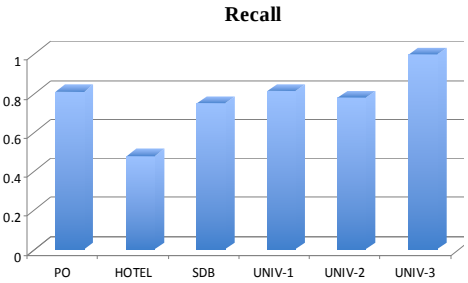

Fig. E.4. Recall values for **SASMINT** - select max above threshold strategy

E.3 Evaluation of Schema Matching Using F-Measure

When precision and recall values are combined using f-measure, SASMINT and COMA++ accomplished almost the same for the hotel, SDB, and UNIV-2 schemas. However, for the remaining schemas, SASMINT performed around 1.1 times better than COMA++. The

average f-measure for SASMINT was 0.84, whereas for COMA++ it was 0.80 and thus the quality of SASMINT's results is better than COMA++. F-measure values for COMA++ and SASMINT over all schema pairs are shown in Figures E.5 and E.6 respectively.

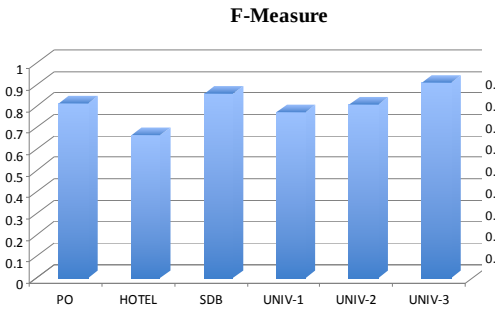

Fig. E.5. F-measure values for **COMA++** - select max above threshold strategy

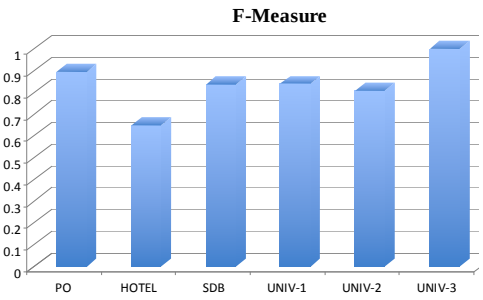

Fig. E.6. F-measure values for **SASMINT** - select max above threshold strategy

E.4 Evaluation of Schema Matching Using Overall

Situation for the overall measure was similar to f-measure, except for the SDB schema. For this pair, the value for COMA++ was slightly (1.05 times) better than SASMINT. For the hotel and UNIV-2 schemas, overall values for SASMINT and COMA++ were the same. However, for the purchase order, UNIV-1, and UNIV-3 schemas, SASMINT performed 1.2 times better than COMA++. The average overall value for SASMINT was 0.72, whereas for COMA++ it was 0.66. Complete results for COMA++ and SASMINT are shown in Figures E.7 and E.8 respectively.

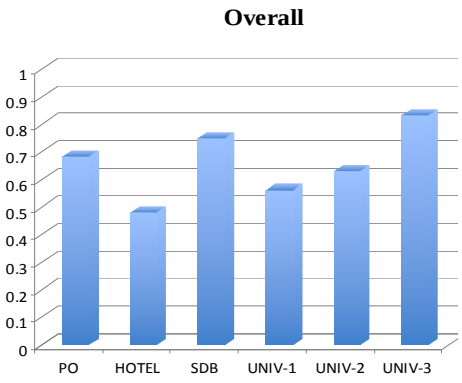

Fig. E.7. Overall values for **COMA++** - select max above threshold strategy

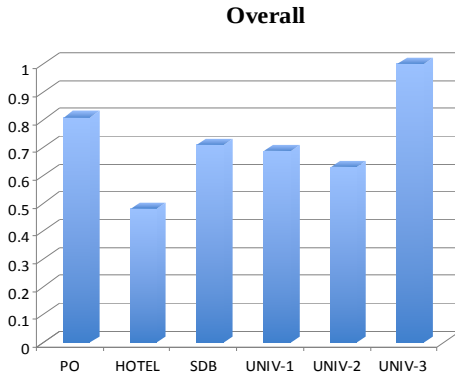

Fig. E.8. Overall values for **SASMINT** - select max above threshold strategy