Supplement for Improving Comprehension of Measurements Using Concrete Re-expression Strategies

COMPARISON TO HUMAN-RANKED RE-EXPRESSIONS

We used psychology research to motivate a ranking of the criteria in our energy functions by importance, then repeatedly assessed results to identify the specific term weights. An alternative method for setting the term weights is to have people rank a large set of re-expressions and then use their preferences to learn optimal term weights. For each measure, we compared our system ranking of re-expressions (i.e., the ranking produced using the term weights reported in the paper) to a ranking of the same results that we generated using term weights that we optimized based on a human-ranked gold standard set. We used crowdsourcing to obtain gold standard ranks for sets of reunitizations at varying reference magnitudes for each measure. We then learned the best weights by searching over a large set of possible weights for those that minimized distance to the crowd ranking. Overall, we find that the rankings produced by our reunitization strategy are comparable to those optimized to the crowd rankings.

Specifically, we used a grid task similar to that used to obtain crowd familiarity scores for each object. However, rather than showing objects only, the grid contained reunitizations. Each reunitization was depicted as an object image and label plus the multiplier (e.g., "2.3 basketballs"). Additionally, rather than asking workers to assign a rank only to the top four most familiar grid elements as in the crowd familiarity task, we instead asked workers to assign a rank to all of the reunitizations that appeared in each grid. We presented a smaller number of reunitizations in each grid (8) to make the task easier for crowdworkers (in the earlier task, workers had to rank the familiarity of only the top four out of 16 objects in the grid). We generated the grids for each measure excluding price (weight, height, length, volume) and reference measurement from the set [0.01, 0.05, 0.1, 0.5, ... 10,000] by randomly sampling 8 reunitization objects from a larger set of 30 objects that were ranked as most familiar by our crowd familiarity score for that reference measurement and measure. We used the most familiar objects to help ensure that the crowd ranking results we obtained could be used to differentiate between reunitizations that used objects of similar, high familiarity. We chose a sample size of 30 as it was likely to be large enough to include the most familiar set of objects despite possible variance in the crowd familiarity rankings. We generated 1,560 total grids, such that for each of the four measures and thirteen inputs, 30 workers ranked a set of eight randomly selected familiar objects. Each HIT carried a reward of $0.20 and presented a worker with a single grid. Each HIT was completed by 30 workers from the Amazon Mechanical Turk Master pool.

Our goal in asking workers to rank reunitizations was to obtain a set of "gold standard" rankings to compare to our automated reunitization strategy's rankings of the same reunitizations. For each measure and reference measurement pair (e.g., weight of 0.01 pounds) we calculated the crowdworkers' aggregate median rank of each object that appeared in at least one grid. We used the median rank to avoid the biasing effect of outliers on mean ranks. For each measure and reference measurement pair, we then calculated the Spearman's rank correlation *rs* for rankings generated by our system and the crowd median rankings. For the measure weight, the mean correlation across all reference measurements was 0.24; for height, 0.37, for length, 0.25, and for volume, 0.25. An alternative way of comparing the alignment between the crowd rankings and our baseline reunitization rankings is to consider how often the rank for a reunitization falls within the range (minimum - maximum) of crowdworker ranks for that reunitization. Over all four measures, our system ranks are within the range of crowdworker ranks for 64% of all ranked reunitizations (by measure: weight: 75%; height: 64%; length: 60%; volume: 61%). We see moderate correlations between our reunitization strategy's rankings and the crowd rankings across the set of measures using both measures of alignment between rankings.

Additionally, we used the gold standard rankings to find the "optimal" term weights for the re- unitization implementation for each measure. For each measure, we searched over a large space of possible weights for w_{omf} , w_{mv}, and w_{mult} in our reunitization function, seeking the set of weights that minimized the distance between the

system rankings and the crowd rankings. We used the mag- nitude of the vector 1-r_s as a distance measure. For each measure, we used this method to evaluate term weights ranging between 0.01 and 1 in increments of 0.05. Using this approach, we arrived at the following optimal weights for the terms [w_{omf}, w_{mult} , w_{mv}] : for weight [0.96, 0.06, 0.36]; for height [0.41, 0.01, 0.16]; for length [0.46, 0.96, 0.01]; for volume [0.46, 0.01, 0.56].

To evaluate how much these optimized term weights improve the ranking of reunitizations relative to the human ranked gold standard, we did several analyses. First, for each reference measurement, we calculated the Spearman's rank correlation *rs* for the optimized rankings and the crowd median rankings. We compare these correlations to those obtained for our baseline reunitization implementation. While the optimized rankings lead to higher r_s for each of the four measures, the gains are relatively small. For the measure weight, the optimized and baseline reunitization correlations with the crowd median rankings are 0.39 and 0.24, respectively; for height, 0.48 and 0.37, for length, 0.36 and 0.25, and for volume, 0.36 and 0.25. Hence, we find that the optimized weights only slightly improve the correlations.

Comparing how often the rank for a reunitization falls within the range of crowdworker ranks for the optimized rankings and our baseline reunitization rankings indicates little difference between the two methods. The optimized ranks are within the range of crowdworker ranks for 64.5% of all ranked reunitizations over all four measures, compared to 64% for our baseline implementation. By measure, the optimized ranks are within the range of crowdworker ranks 73% for weight (compared to 75% for our baseline implementation); 65% for height (compared to 64%); 62% for length (compared to 60%); and 58% for volume (compared to 61%).

We conclude from these analyses that optimizing the term weights in our automated implementations can slightly improve the degree of agreement between human rankings and the automated implementation. However, these gains are small, most likely as a result of the natural variance between users when asked to rank reunitizations.

Relative Importance of Energy Terms

We can also assess the relative importance of each term in the reunitization energy function by comparing the distances and average correlations between the crowd rankings and the ranked results generated with all three terms in the energy function versus with single terms left out. We again use the magnitude of the vector 1-r_s as a measure of the overall distance between pairs of rankings, where a smaller magnitude means greater alignment between the two sets of rankings. Comparing the magnitude of this vector across all measures for input measurements in the set [0.01, 0.05, 0.1, 0.5, ... 10,000], we find that the mean magnitude is 2.78 for all terms, 2.97 when we leave out the familiarity term, 2.88 when we leave out the measure vari- ance term, and 3.02 when we leave out the multiplier term. Across all measures, this corresponds to a smaller distance with the crowd rankings by an average of 6.4% for all terms versus leaving out the familiarity term, by 3.5% for all terms versus leaving out measure variance term, and by 8.0% for all terms versus leaving out the multiplier term.

Similarly, comparing correlations across all measures, we find that the mean correlation with the crowd rankings across input measurements in the set [0.01, 0.05, 0.1, 0.5, ... 10,000] is 0.27 for all terms, 0.24 when we leave out the familiarity term, 0.26 when we leave out the measure variance term, and 0.22 when we leave out the multiplier term. This corresponds to an improvement in the mean correlation across all measures of 13% for all terms versus leaving out the familiarity term, of 4% for all terms versus leaving out the measure variance term, and of 23% for all terms versus leaving out the multiplier term. Though improvements in both the vector magnitude and correlations are relatively small, we see a consistent improvement with the addition of each term in the energy function, with the greatest improvements coming from the familiarity and multiplier terms, which we weight higher than the measure variance term. Results for each reference measurement and measure are available as supplemental material.

We also provide a sample of the leave-one-term-out results for a set of 15 input measurements to allow for more qualitative analysis of how the terms influence the rankings. The supplementary file **leave_out_term_comparison_results.xlsx** is an Excel workbook containing analyses results from leaving one term out of the reunitization energy function at a time. The first sheet shows the top three reunitizations returned by our solution for the same set of input measurements, either using all three terms (object-measure familiarity, measure variance, and multiplier, labeled FULL) in our energy terms or leaving one term out of the energy function at a time. The second through fifth sheets compare our reunitization solution with all three terms in the

energy function (FULL) to those that leave out one term at a time in terms of distance to and correlation with the crowd-based gold standard rankings for the set of input measurements [0.01 – 10000]. The columns *vectorMag* and *vectorSum* are measures of the distance between the two pairs of rankings. The columns *cor0.01* - *cor10000* are the correlations between the crowd rankings and the given ranking. The columm *meanCorr* is the average correlation between the two pairs of rankings across all input measurements. The columns *famweight*, *multweight*, and *varweight* are the weights of the object measure familiarity term, the multiplier term, and the variance term, respectively.