Enhancing Protocol Evaluation Through Semantic Modification of Local Benchmarks

Shravan Mylavarapu[1], Irwin Levinstein[2], Chutima Boonthum[3], Joseph Magliano[4], Keith Millis[4]

[1] Fidelity Investments, USA
[2] Old Dominion University, Norfolk, VA, USA
[3] Hampton University, Hampton, VA, USA
[4] Northern Illinois University, DeKalb, IL, USA
{shravanmsv@yahoo.com, ibl@cs.odu.edu, chutima.boonthum@hamptonu.edu, jmagliano@niu.edu, kmillis@niu.edu}

Abstract

A computer based evaluation is essential to allow us to assess students' protocols without human intervention in R SAT (Reading Strategy Assessment Tool). Word matching and Latent Semantic Analysis (LSA) approaches were explored for use in the evaluation of protocols and. experimentally it was found that word matching alone outperformed LSA or LSA in combination with word matching. *Why does LSA not work well with the R SAT protocol evaluation?* This paper demonstrates that modifying the benchmarks on which it is based can indeed benefit its performance. A variety of local benchmark modifications are investigated and results are compared against human expert. The preliminary results show that modified local benchmarks improve the R SAT protocol evaluation on *local bridging* strategy: 0.1 increase in correlation and 8% increase in percent agreement between original benchmarks and modified benchmarks.

Introduction

R-SAT (Gilliam et al., 2007) is an automated assessment tool for identifying weaknesses in students' reading comprehension strategies. In the course of its development three strategy-evaluation techniques have been considered: (1) Human Coding by a human expert to provide a basis for comparison for the following automated techniques; (2) Word Matching technique, based on partial prefix matching of student content with text content; and (3) Semantic Matching technique, based on conceptual matching of student content with text content.

Since the word-matching algorithm is more successful in R-SAT protocol evaluation, we decided to investigate ways to reduce the over generosity of the LSA algorithm. This led to the idea of a reduction of overlap among benchmarks. In a previous unpublished analysis, a slight improvement was gained when overlap was reduced by *deleting duplicate words* (literal match) from one of benchmarks. In the new approach, a word is to be removed from a benchmark when it *semantically* contributes to an overlap of two benchmarks. This approach seeks to avoid the same protocol content to be given credit in both benchmarks.

For each reading strategy, a standard *benchmark* is predefined so that it can be used to compare against the student's protocol. The benchmark is simply a "bag" of content words, i.e., order is not significant. The *current sentence* is a benchmark for *paraphrasing*; *immediate prior sentence* is for *local bridging*, *distal prior sentences* for *distal bridge*, and *subsequent sentences* is for *elaboration*.

Overlap Reduction

To make the LSA algorithm behave more like the word-based algorithm, the semantic overlap between benchmarks were systematically removed. In this approach, a word is removed from a benchmark when it *semantically* contributes to an overlap of two benchmarks. This approach seeks to avoid allowing the same protocol content to be given credit in comparison with both benchmarks.

The aim of this research was to discover the best strategy for reducing semantic overlap between benchmarks. The plan was as follows:

1. Identify the words that are the cause for the semantic overlap; *high impact words*.
2. Sort the high impact words from highest to lowest impact.

3. Apply different strategies for removing these words (A. Always remove *n* high impact words; B. Remove up to *n* high impact words based on a *threshold* of permissible overlap; or C. Remove up to *n%* high impact words based on a *threshold* of permissible overlap)
4. Determine which strategy yields revised protocols that produce the best prediction of the human evaluation of the strategy used in the protocols.

Results

Correlation, percent agreement, and d' (d-prime) between the proposed method (semantic modification of benchmarks) and human coding were calculated to show the performance. The baseline of this analysis is the original word matching algorithm's result where overlap between benchmarks is present. Four different base data sets were used while experimenting. In the table below, d' 0 measures absence of a strategy; d' 1, partial presence and d' 2, complete presence.

Method	R	%	d' 0	d' 1	d' 2
ORG	0.345	67.7	1.124	0.590	0.810
R1	0.356	75.3	1.182	0.486	0.855
R2	0.335	76.3	1.219	0.413	0.767
R3	0.339	76.7	1.294	0.345	0.738
R3T1	0.374	75.9	1.213	0.551	0.935
R3T2	0.400	76.1	1.302	0.629	0.963
R3T3	0.433	76.0	1.451	0.741	0.918
R3T4	0.450	75.3	1.480	0.752	0.976

Table 1: Mechanically chosen local benchmarks with overlap

ORG Original result from word based algorithm (no overlap)
R1 Always remove 1 word
R2 Always remove 2 words
R3 Always remove 3 words
R3T1 Remove up to 3 words using threshold 0.1
R3T2 Remove up to 3 words using threshold 0.2
R3T3 Remove up to 3 words using threshold 0.3
R3T4 Remove up to 3 words using threshold 0.4
R Correlation compare to human coding
% Percent agreement against human coding

Method	R	%	d' 0	d' 1	d' 2
ORG	0.310	68.5	1.086	0.704	0.608
R1	0.319	75.4	1.178	0.702	0.570
R2	0.313	76.1	1.190	0.592	0.592
R3	0.318	76.5	1.242	0.544	0.573
R3T1	0.351	75.6	1.200	0.627	0.702
R3T2	0.373	75.8	1.280	0.732	0.708
R3T3	0.407	75.9	1.430	0.838	0.686
R3T4	0.425	75.5	1.484	0.914	0.761

Table 2: Mechanically chosen local benchmarks no literal overlap

The correlation and d' values appeared satisfying although the % agreement was a little low compared to *chosen local benchmarks no literal overlap.* Now, looking at the strategy that gave satisfactory results, we find that *Remove up to 3 words using threshold 0.4* appears to be the best because of high d' values and high correlation.

To verify the result, the data was split into two sets: *training set* and *test set*. The predicted formulae are obtained from the training set and applied on the test set.

Method	R	%	d' 0	d' 1	d' 2
Train	0.318	67.7	1.084	0.803	0.770
Test	0.300	67.6	1.259	0.749	0.437
Overall	0.306	67.6	1.110	0.765	0.563
R3T4: Train	0.510	78.1	1.640	0.871	1.027
R3T4: Test	0.384	74.8	1.428	0.949	0.500
R3T4: Overall	0.421	75.8	1.485	0.923	0.667

Table 3: Mechanically chosen local benchmarks with overlap: split data

This result confirms that removing up to 3 words using threshold 0.4 is the best benchmark modification method for local benchmark. Focusing only presence (1 and 2 combined) or absence (only 0) of such strategy is also significant.

Method	R	%	d' 0	d' 2
R3T4	0.495	82.6	2.747	2.747

Table 4: Mechanically chosen local benchmarks with overlap: focusing on absence and presence of a local bringing strategy.

Conclusion

Modifying local benchmarks by removing the high impact word showed a significant improvement in the R-SAT evaluation. With this, we are able to identify whether the student has used *local bridging* strategy in their input. This is a good indication that LSA definitely benefit and can be used for R-SAT strategy evaluation.

References

Gilliam, S., Magliano, J.P., Millis, K.K., Levinstein, I.B., and Boonthum, C. 2007. Assessing the format of the presentation of text in developing a Reading Strategy Assessment Tool (R-SAT). *Behavior Research Methods Instruments and Computer*, 39 (20), 199-204.

Landauer, T.K., McNamara, D.S., Dennis, S., and Kintsch, W. eds. 2007. *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum, Mahwah, NJ.

McNamara, D.S., Boonthum, C., Levinstein, I.B., and Millis, K.K. 2007. Evaluating self-explanation in iSTART: Comparing word-based LSA systems. T. Landauer, D.S. McNamara, S. Dennis, and W. Kintsch eds., *Handbook of Latent Semantic Analysis* (pp. 227-241). Lawrence Erlbaum, Mahwah, NJ.