Using Latent Semantic Analysis and Word Matching to Enhance Bridging Reading Strategy Identification

Martha Brhane and Chutima Boonthum-Denecke
Department of Computer Science, Hampton University, Hampton, VA 23668
martinameskel@yahoo.com, chutima.boonthum@hamptonu.edu

Abstract

The main goal of this study is to identify bridging reading strategy -- a strategy that a reader uses to make a connection from the current sentence to previous sentences to help understanding the meaning of the text. For a specific target sentence, there are two types of bridging: local and distal. Benchmarks were created to help represent each type of bridging. The two immediate prior sentences of each target sentence together created a benchmark for the local bridging. The benchmarks for distal bridging were those prior sentences, excluding two immediate prior sentences. There were three ways that distal benchmarks were created: chunks based-on paragraph, chunks based-on target sentence, and entire collection of prior sentences. The results showed that using modified benchmark by removing up to 4 words within a threshold 0.4 has significantly improved the identification of distal bridging reading strategy by 14% from the original benchmark evaluation. On the other hand, to identify local bridging, using modified benchmark by removing 4 words has significantly improved the identification by 19% from the original benchmark evaluation.

Introduction

Reading strategies such as comprehension monitoring, paraphrasing, elaboration, prediction, and bridging can help struggling adolescent readers to build the skills they need to succeed in high school and beyond. iSTART (Interactive Strategy Training for Active Reading and Thinking) is a web-based application that provides students with self-explanation and reading strategy training. (McNamara, Levinstein, and Boonthum 2004).

Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a Natural Language Processing technique for determining the similarity of meaning of words and passages by analysis of large text corpora (Landauer and Dumais 1997).

For any pair of documents, the semantic similarity can be compared by calculating the cosine measure between their document vectors as

$$Sim(D1, D2) = \frac{\sum_{i=1}^{d}(D1_i X\ D2_i)}{\sum_{i=1}^{d}(D1_i)^2\ X\ \sum_{i=1}^{d}(D2_i)^2}$$

The value of LSA cosine value is from -1 to 1. A higher cosine indicates that two units of text are more semantically similar (McNamara et al. 2004).

Word Matching

Word matching is another Natural Language Processing technique in which it compares character by character of words for evaluating natural language. Word matching can be performed in two ways: (1) Literal word matching and (2) Soundex matching.

Benchmarks

Benchmark is bag of words that represent each of the different reading strategies. A target sentence is a sentence has a high conceptual connection with other prior sentence of the text (Malladi et al. 2010).

Bridging Reading Strategy

Bridging is one of the reading strategies in which a reader explains the current sentence using concepts that are previously mentioned in the text. Bridging can be local or distal (Gilliam et al. 2007).

Constructing Bridging Benchmarks

Local Bridging Benchmark - the two immediate prior sentences of each target sentence.

Distal Bridging Benchmark - those prior sentences of each target sentence, excluding the two immediate prior sentences. Distal bridging benchmarks were derived in

three different chunking ways: (1) chunks based-on paragraph, (2) chunks based-on target sentence, and (3) entire collection of prior sentence.

Versions of Bridging Benchmarks

For each target sentence there are two versions of benchmark: **Original benchmark** - which created by removing stop words. **Modified benchmark** - which created by removing overlapping words between the original benchmarks. Overlapping words were removed using literally and semantically approaches.

(1) Removing Overlapping Words Literally - if there was 70% or 80% match between words of the two documents whose word length is greater than five, then the word was removed from the original benchmark, otherwise an exact word match was required for word whose length was less than or equal to five.

(2) Removing Overlapping Words Semantically - words that contributes semantically the same meaning between benchmarks were removed in three ways: (1) removing 3, 4, 5, and 6 high impact words, (2) removing up to 3, 4, 5, and 6 words within threshold of 0.3, 0.4, 0.5, and 0.6, and (3) removing up to 30%, 35%, and 40% words within threshold of 0.3, 0.4, 0.5, and 0.6.

Bridging Reading Identification

This study was conducted on the explanation of sixty one (61) students that was obtained from R-SAT participants at Northern Illinois University (Gilliam et al. 2007). Each explanation was rated by a human expert as 0, 1, or 2 which means bridging not used, partially used, or fully used respectively. The data obtained for the students were divided into two sections: training set - the first half section and was used to create formulae. The rest half section was defined as test set and was used to validate the consistency of the formulae. Logical expression, logical binomial expression, regression analysis, and discriminant analysis were the different formulae constructed to identify students' explanation for bridging reading strategy.

Results and Conclusion

Correlation and percent agreement were computed among results obtained using different formulae and human expert scores. Results showed that paragraph chunk benchmark type gave a better result comparing with other type of benchmarks. Figure 1 explains removing four words within threshold of 0.4 gives relatively a better result in all the formulae than using the rest of the threshold.

Figure 2 shows using logical binomial formula removing four words improves the original benchmark for local bridging.

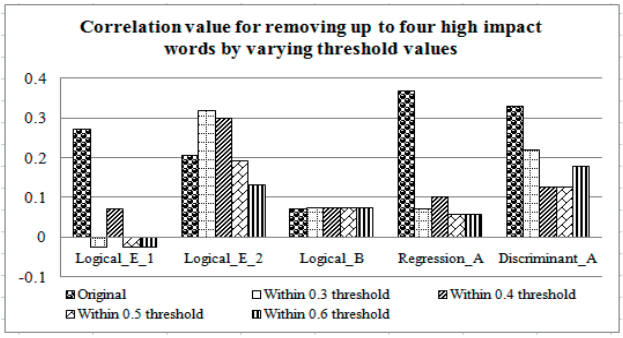

Figure 1. Correlation for removing up to four words by varying threshold (paragraph chunk).

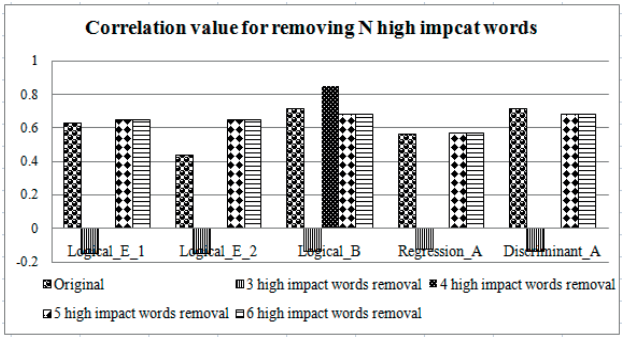

Figure 2. Correlation for removing N words (local bridging).

In summary, from this study, for distal bridging benchmark removing up to four words within threshold of 0.4 improved the strategy identification by 14%. On the other hand, the use of modified benchmark by removing four words to identify local bridging reading strategy has significantly improved the strategy identification by 19%.

References

Gilliam, S., Magliano, J. P., Millis, K. K., Levinstein, I., and Boonthum, C. (2007). Assessing the format of the presentation of text in developing a reading strategy assessment tool (R-SAT). *Behavior Research Methods*, *39*(2), 34-44.

Landauer, T. K., and Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, *104*(2), 211-240.

Malladi, R., Levinstein, I., Boonthum, C., and Magliano, J. (2010). Summarization: Constructing an ideal summary and evaluating a student's summary using LSA. In *Proceedings of the 23rd FLAIRS Conference* (pp. 295-296). Menlo Park, CA: The AAAI Press.

McNamara, D. S., Levinstein, I. B., and Boonthum, C. (2004). iSTART: Interactive strategy training for active reading and thinking. *Behavior Research Methods, Instruments, and Computers, 36*(2), 222-233.