SEAS RESEARCH & DEVELOPMENT SHOWCASE

Automatic Generation of Context-Based Fill-in-the-Blank Exercises Using Vector Space Models and Google n-grams

Jennifer Hill (Advisor: Rahul Simha)

INTRODUCTION
According to the American Library Association, 14% of adults in the United States cannot “search, comprehend, and use continuous texts” [1]. Current government and philanthropic funding only indirectly helps one tenth of the nearly 30 million individuals.

To address this shortcoming, we aim to automatically create reading comprehension exercises from existing text passages. We specify that a successful comprehension exercise should challenge a reader’s contextual understanding of the passage’s meaning rather than his or her vocabulary understanding. We propose a method of automatically generating fill-in-the-blank exercises to challenge and improve comprehension skills, using a unique application of word co-occurrences vector space models and the Google n-grams database.

CHOOSING BLANKS
We choose to blank out words that have strong contextual links to words in the surrounding text, leaving enough context for the reader to understand the sentence’s intended meaning when that word is removed.

To determine contextually-linked words, we utilize word co-occurrence likelihoods from the word vector space model GloVe [2]. We assume that words that are paired together regularly are likely to have a notable contextual and semantic relationship.

CHOOSING DISTRACTORS
We explore a unique application of the Google n-grams corpus [3] for generating distractors for our fill-in-the-blank questions. We find all words with the same part of speech as the blanked word found in the Google n-grams database.

EVALUATION
Our corpus contained 18 passages obtained from ReadWorks.org (Lexile Level 100 to 1000). For each passage, we generated fill-in-the-blank questions for each scope, resulting in 170 unique questions.

53 human volunteers answered an anonymous questionnaire with both blanked phrases and full passage exercises. We evaluate our questions on their validity and reliability.

RESULTS & CONCLUSIONS
We choose to blank out words that have strong contextual links to words in the surrounding text, leaving enough context for the reader to understand the sentence’s intended meaning when that word is removed.

Algorithm for selecting blanked words by finding closest co-occurring pairs.

Algorithm for selecting distractors by finding matching n-grams matches.

Table 1: Distribution of ratings for all generated questions

Table 2: Percentage of target and distractor words determined to fit each blank

EXAMPLE

"It's written for and put together by the fifth graders," Dr. Reed said. "It’s written for and put [blank] by the fifth graders," [PRP] said.

Validly
- Blanks: rate question on a 1-5 quality scale (without distractors)
- Distractors: examine ratio of words that fit blank in narrow context to words that fit broader context.
  - Narrow: Correct 100%, Incorrect 0%
  - Full: Correct 100%, Incorrect 0%

Reliability
- Distractors: examine percentage of correct answers selected when given the full context

REFERENCES