ISCA Archive
http://www.isca-speech.org/archive

Speech and Language Technology
in Education (SLaTE2007)
Farmington, PA, USA
October 1-3, 2007

SourceFinder: A Construct-Driven Approach for Locating Appropriately Targeted Reading Comprehension Source Texts

Kathleen M. Sheehan, Irene Kostin, Yoko Futagi

Educational Testing Service
Princeton, NJ, USA
{ksheehan,ikostin,yfutagi}@ets.org

Abstract

A fully-automated approach for locating source material for use in developing reading comprehension/verbal reasoning passages is described. The system employs a combination of classification and regression techniques to predict the acceptability status of candidate source texts downloaded from targeted on-line journals and magazines. The approach is applied to the problem of selecting source texts pitched at a particularly advanced reading level, i.e., the level expected for students seeking admission to graduate school. Results confirm that, even at this advanced level, SourceFinder behaves much like a human rater. In particular, while the human raters agreed with each other 63% of the time, the agreement between SourceFinder and a human rater ranged from 61% to 62%. This suggests that the estimated models have succeeded in capturing useful information about the characteristics of texts that affect test developers' ratings of source acceptability and that continued use of the system may help test developers find more high quality sources in less time.

1. Introduction

New test delivery technologies, such as Internet-based testing, have created a demand for higher capacity item generation techniques that are (a) grounded in a credible theory of domain proficiency, and (b) aligned with targeted difficulty specifications. Since many existing testing programs employ stimulus passages that have been adapted from previously published source texts, researchers at the Educational Testing Service (ETS) have developed an automated text analysis system designed to help test developers locate appropriately targeted stimulus materials more efficiently. This new system, called SourceFinder, includes three main components: (a) a corpus of candidate source documents downloaded from targeted online journals and magazines, (b) a source evaluation module (SEM), and (c) a capability for efficiently searching the corpus so that users (i.e., test developers) can restrict their attention to only those sources that have been rated as having a relatively high probability of being acceptable for use in the particular source-finding assignment at hand. Since stimulus requirements vary considerably both across and within testing programs, the SEM evaluates each candidate source document multiple times. Results are then communicated to users via a set of acceptability ratings defined such that each individual rating reflects the acceptance criteria appropriate for a specific type of passage associated with a specific testing program.

This study reports results obtained for the paragraph reading (PR) item type, a new item type developed for use on the Graduate Record Examination (GRE), an examination taken by students seeking admission to graduate school. The PR item type consists of a short passage followed by two to four items designed to elicit evidence about an examinee's ability to understand and critique complex verbal arguments such as those that are typically presented in scholarly articles targeted at professional researchers. This new item type was developed at ETS as part of an on-going effort to enhance the validity, security and efficiency of item development procedures for the GRE.

The remainder of this paper is organized as follows. Relevant prior work is reviewed in Section 2. Our training and validation corpora are described in Section 3. The text features considered in the analyses are described next (Section 4), followed by a description of the modeling technique (Section 5) and validity evidence (Section 6). A final section summarizes conclusions and directions for future research (Section 7).

2. Previous work

Previous reading level assessment research is described in [1] and [2]. The current study differs from this earlier work in terms of the specific reading level assessed and the estimation techniques employed. While much of the previous research has been focused at the K-12 level, the current application considers the reading level expected for students seeking admission to graduate school. Also, while the previous research considered large numbers of independent variables evaluated via machine learning techniques (e.g., Naïve Bayes, k-Nearest Neighbor algorithms and Support Vector Machines) SourceFinder considers a somewhat smaller set of independent variables designed to provide enhanced stability, while still maintaining construct relevance and ease of interpretation. These are developed via two techniques: (a) a corpus-based dimensionality analysis similar to that described in [3] and [4], and (b) content vector analyses similar to those described in [5], [6] and [7].

3. Corpus development

Training and validation corpora were developed as follows. First, an initial training sample was assembled by randomly selecting 114 paragraphs from a database of targeted on-line journals and magazines. The selected paragraphs were then presented to two GRE test developers for evaluation. Raters provided two types of evaluations: (a) a quantitative estimate of the paragraph's "acceptability" status expressed on a 1 to 5 scale, where 1 = *definitely reject*, 2 = *probably reject*, 3 = *uncertain*, 4 = *probably accept*, and 5 = *definitely accept*, and (b) a brief, written description of the aspects of text variation considered during the rating process. Because the resulting

sample was not expected to yield a large number of acceptable paragraphs, a supplemental training sample of 47 *historical* paragraphs was also assembled. Historical paragraphs are paragraphs that had previously been used to create operational PR passages. This strategy yielded 47 additional training paragraphs classified at the *definitely accept* level, and increased the size of the training sample to a total of 161 paragraphs.

An independent validation sample was also assembled. It consisted of 1,000 additional paragraphs selected from the same database. Unlike the initial training sample (which had been sampled randomly) some of the validation paragraphs were selected via a nonrandom process, e.g., searching favored journals first. As was the case for the training sample, however, each paragraph was independently rated by two test developers. A total of 14 experienced test developers participated in the rating process.

4. Features

Each rater provided a brief, written description of the individual text characteristics considered while rating each paragraph. These comments constituted the primary data considered during feature development. The comments suggested that, at a minimum, the raters tended to focus on three particular aspects of text variation: (a) rhetorical style, (b) content, and (c) sensitivity. The following paragraphs describe the individual text features developed to characterize text standing relative to these aspects

4.1 Rhetorical style

Each GRE passage must be capable of supporting the types of complex reasoning items needed to provide accurate measurement at the high end of the GRE scale. Texts that are primarily descriptive or that merely present straight-forward exposition or narration are less likely to support challenging reasoning items, while texts that provide some conflict or contrast of ideas and some uncertainty about conclusions or outcomes are more likely to support such items. The following comments were judged to be indicative of a violation relative to this particular aspect of text variation: (a) "*Not enough tension/argument*." (b) "*Not really any reasoning here*," and (c) "*Too thin. Descriptive rather than reasoning*."

A total of 42 different text features were developed to quantify potentially informative aspects of rhetorical style. Many of these were based on previous research documented in [3] and [4]. Because this set was so large, and because many of the resulting features were highly correlated, a factor analysis (FA) was used to define linear combinations of features for use in model development. The strategy of using FA to explore the patterns of linguistic variation detected in representative collections of texts is discussed in [3] and [4]. For example, [4] argues that, because many important text characteristics are not well captured by individual linguistic features, investigation of such characteristics requires a focus on "constellations of co-occurring linguistic features" as opposed to individual features. FA permits easy access to such "constellations" by allowing patterns of linguistic co-occurrence to be analyzed in terms of underlying "dimensions of variation" or "factors" that are identified quantitatively.

In order to provide a more stable solution, the FA was implemented with respect to entire documents, as opposed to individual paragraphs. A total of 937 documents sampled from the database of on-line journals and magazines was selected for use in the analyses. Each document contained between 1,000 and 5,000 words, yielding a total corpus size of more than 4.5 million words. The major dimensions of variation underlying the candidate features were identified by implementing a principal component analysis extraction followed by a Promax rotation. A principal component analysis extraction was selected because our primary goal involved reducing a large number of candidate features down to a more manageable number of dimension scores. A Promax rotation was selected because the resulting dimension scores were expected to be moderately correlated.

The results suggested that, at most, eight dimensions of variation were being measured. The eigenvalues for these eight components were as follows: 10.473, 4.889, 2.741, 2.132, 1.879, 1.557, 1.340 and 1.1.30. Since only the first six factors appeared to be construct relevant, a six-factor solution was extracted. Taken together, these six factors accounted for nearly 60% of the shared variance.

Table 1 lists illustrative features and loadings for these six dimensions. The Table also provides a short descriptive label for each dimension. These were developed by considering the pattern of variation implied by the highly weighted features within each dimension. The results suggest that the rhetorical style of candidate GRE source texts can be decomposed into the following six dimensions of variation: (1) degree of narrative orientation; (2) degree of academic orientation; (3) amount of overt argumentation; (4) amount of opposition; (5) sentence complexity, and (6) vocabulary level.

This solution yielded six candidate explanatory variables for consideration in the model development activities described below. Each variable was defined as a linear combination of 42 text features, with coefficients selected to approximate the document-to-document correlation structure.

Table 1. *Dimensions of variation with illustrative features and loadings*

Dimension/Feature	Loading
1: Narrative Orientation	
Communication Verbs (e.g., *say,, call*)	+.84
Third Person Singular Pronouns (e.g., *he, she*, etc.)	+.53
2: Academic Orientation	
Cognitive process/perception nouns (e.g., *concept*)	+.96
Abstract Concept Nouns (e.g., *existence, progress*)	+.75
3: Overt Expression of Argumentation	
Possibility modals (e.g., *can, can't, could*, etc.)	+.77
Prediction modals (e.g., *shall, will, won't*, etc.)	+.61
4: Opposition	
Analytic Negation (e.g., *not*)	+.67
Oppositional reasoning words (e.g., *challenge*, etc.)	+.54
5: Sentence Complexity	
Median Length of Longest Clause (log words)	+.88
Average Sentence Length (log words)	+.87
6: Vocabulary	
Word Types Not in the EWFG	+.78
Average EWFG Word Frequency	-.63

Notes. EWFG = The Educators Word Frequency Guide, a word frequency index published by Touchstone Applied Sciences Associates in 1995. Except where noted, all features are expressed on a log frequency per thousand words scale.

4.2 Content

Content features were developed by using a content vector analysis ([5], [6]) to quantify the degree of similarity between candidate source documents and a set of target content vectors constructed to characterize vocabulary usage within the four main GRE content areas, i.e., physical sciences (PS), biological sciences (BS), social sciences (SS) and humanities (HU). The analysis was implemented as follows. First, a set of 261 previously administered GRE passages with known content classifications was used to construct four target content vectors, one for each of the four areas listed above. The vectors provided content-area specific word frequency information for the set of all nouns, verbs, adjectives and adverbs that, after stemming, appeared in at least 2 of the 261 passages. Because the resulting vectors were relatively sparse, a similarity index [8] was then used to collapse across rows indexed by similar content words. For example, *bacteria* and *germ* were collapsed because their similarity score fell above an empirically determined threshold. The resulting condensed vectors were then used to estimate four cosine scores for each document, one for each of the four GRE content areas. These scores were then treated as additional explanatory variables.

Portions of the resulting target vectors are shown in Table 2. The vectors suggest that the four main GRE content areas tend to employ relatively distinct vocabularies. For example, words like *species*, *population*, *brain*, *process*, *bacteria* and *germ* tend to occur with relatively high frequency in biological science passages and relatively low frequency in each of the other three types of passages. Similarly, words like *art*, *work*, *literary*, *artistic*, *writer*, and *novel* tend to occur with relatively high frequency in humanities passages and relatively low frequency in each of the other three types of passages.

Table 2. *Standardized term frequencies for selected word classes (in Frequency per 1000 words)*

Word Class	**BS**	**PS**	**SS**	**HU**
Species (N)	**3.69**	0.33	0.15	0.03
Population (N)	**2.78**	0.00	0.93	0.00
Brain (N)	**2.01**	0.00	0.00	0.00
Process (V)	**1.75**	0.92	0.06	0.21
Bacteria|Germ (N)	**1.49**	0.00	0.00	0.00
Surface (N)	0.19	**4.29**	0.03	0.18
Earth (N)	0.39	**4.09**	0.00	0.07
Star (N)	0.00	**3.83**	0.00	0.00
Planet (N)	0.00	**3.10**	0.00	0.00
Electron|Neutron|Particle (N)	0.00	**1.91**	0.00	0.07
Political|Ideological (A)	0.06	0.13	**3.06**	0.57
Societal|Social (A)	0.00	0.00	**3.00**	0.75
Historian (N)	0.00	0.00	**2.85**	0.50
Class (N)	0.00	0.00	**1.86**	0.64
Movement (N)	0.06	0.13	**1.74**	0.57
Art (N)	0.00	0.00	0.03	**4.41**
Work (N)	0.00	0.20	2.13	**3.06**
Literary|Artistic (A)	0.00	0.00	0.09	**2.88**
Writer (N)	0.00	0.00	0.09	**2.74**
Novel (N)	0.06	0.00	0.00	**2.49**

Notes. Letters in parentheses indicate part of speech as follows: A = Adjective or Adverb, N = Noun, V = Verb. The construction word1|word2 indicates pairs of words that were collapsed via the similarity index [8]. Values highlighted in bold script are row-wise maximums.

4.3 Sensitivity

Additional variables designed to assess text standing relative to the sensitivity aspect of text acceptability were also developed. These were based on an existing list of potentially inflammatory words (e.g., *abortion*, *amputated*, *addicted* and *depressed*, etc.).

5. Model development

The model development phase of the analysis was designed to generate predictions of text acceptability that closely reflected the ratings provided by the GRE test developers. Two types of models were implemented to achieve this goal: a regression model and a filtering model. The regression model was obtained by regressing the test developers' judgments of source acceptability (expressed on the 5-point scale) on the text variables described above, as shown in Equation 1:

$$y_i = \beta_0 + \sum_{k=1}^{K} \beta_k X_{ik} + \varepsilon_i \qquad (1)$$

where y_i is the average acceptability rating obtained for the i^{th} training paragraph, the X_{ik} are the explanatory variables described above, the β_k are coefficients estimated from the available training data and the error terms, ε_i, are assumed to be independently and identically distributed with mean 0 and common variance σ^2.

One limitation of this estimation approach is that it is not designed to detect violations that are serious, yet not observed in the training sample. This limitation was addressed by also implementing a preliminary filtering step as follows. First, the training data were used to establish an acceptability range for a subset of key features. Next, paragraphs with feature values falling outside of the specified ranges were assigned predicted acceptability scores of 1 (*definitely reject*). For example, because PR passages typically vary between 90 and 130 words, the acceptability range for the paragraph length feature was specified as the interval from 50 to 200 words, and all paragraphs falling outside of that range were assigned predicted acceptability ratings of 1 (*definitely reject*).

The above two approaches, i.e., the regression model and the filtering model, were then used to generate a predicted acceptability rating, expressed on the 5-point scale, for each paragraph in the independent validation sample. Validity evidence developed from these estimates is summarized below.

6. Evaluation

Three types of evaluations were implemented. First, we compared the agreement between SourceFinder and a human rater to that between two human raters. Because very few paragraphs were rated at Levels 2 and 4, the analysis was implemented after first collapsing Levels 1 and 2 to form a single *Reject* category, and also collapsing Levels 4 and 5 to form a single *Accept* category. The resulting agreement data are summarized in Table 3.

Table 3 confirms that both SourceFinder and the human raters rejected a large percentage of the candidate paragraphs. In particular, 681 (539+81+61) of the 1000 paragraphs were rejected at the time of the first human rating, 677 (539+80+58) were rejected at the time of the second human rating, and 654 (514+76+64) were rejected by SourceFinder. The table also shows that, overall, the agreement between SourceFinder and

the human raters (measured on the 3-point collapsed scale) was very similar to that exhibited by the human raters. That is, while the human raters agreed with each other 63% of the time, the agreement between SourceFinder and the human raters ranged from 61% to 62%.

Table 3.*SourceFinder/Human Agreement for 1000 Validation Paragraphs*

R1 (rows) vs. R2 (columns)			R1 (rows) vs. SF (columns)			R2 (rows) vs. SF (columns)		
Rej	Un	Ac	Rej	Un	Ac	Rej	Un	Ac
539	81	61	514	83	84	514	74	89
80	31	43	76	31	47	81	26	50
58	45	62	64	23	78	59	37	70

Notes. Row 1 = Reject, Row 2 = Uncertain, Row 3 = Accept
R1 = 1st Human Rating, R2 = 2nd Human Rating,
SF = SourceFinder, Rej= Reject, Un=Uncertain, Ac=Accept

Table 3 suggests that both SourceFinder and the human raters were more precise at evaluating unacceptable sources than at evaluating acceptable sources. To further investigate this, Table 4 lists precision and recall rates for the three different types of paragraphs. The table shows that: (a) when we restrict our attention to only those paragraphs that were *rejected* at the time of the 1st human rating, SourceFinder's precision is 0.79, while human-to-human precision is only slightly higher at 0.80; and (b) when we restrict our attention to only those paragraphs that were *accepted* at the time of the first human rating, both SourceFinder and the human raters achieved a precision level of 0.37. This confirms that SourceFinder was successful at replicating human patterns of precision.

Table 4. *An Analysis of Precision and Recall, Relative to the 1st Human Rating, for 1,000 Validation Paragraphs*

	2nd Rating			SourceFinder		
1st Rating	R	P	F1	R	P	F1
Reject	0.79	0.80	0.80	0.75	0.79	0.77
Uncertain	0.20	0.20	0.20	0.20	0.23	0.21
Accept	0.38	0.37	0.37	0.47	0.37	0.41

Notes. R=Recall, P=Precision, F1=2RP/(R+P)

The practical significance of these results can be appreciated by comparing the percent of acceptable sources located with and without access to SourceFinder. Data relevant to this comparison are summarized in Table 5. The table suggests that the test developers can increase their acceptance rates from the current level of about 16% to between 33% and 37% simply by restricting their attention to only those documents that SourceFinder classifies at the *Accept* level.

Table 5. *Acceptance Rates Calculated With and Without Access to SourceFinder*

Access to SF?	Ra-tings	No. of Docs. Searched	No. of Accepts	Acceptance Rate
No	R1	1000	165	16.5
No	R2	1000	166	16.6
Yes	R1	209	78	37.3
Yes	R2	209	70	33.4

Notes. Docs=Documents. When Access to SF = Yes, only those documents rated at the *Accept* level are searched.

7. Conclusions and future work

This study described the development and validation of SourceFinder, a fully automated approach for helping test developers locate acceptable source material for use in developing new reading comprehension/verbal reasoning passages. The evaluation considered an application focused at a particularly advanced reading level, i.e., the level expected for students seeking admission to graduate school. This application is unusual since, in many previous studies, the focus has been on K-12 level texts. The results confirmed that, even at this advanced level, SourceFinder's predictions of source acceptability are very similar to those generated by human raters. That is, while the human raters agreed with each other 63% of the time, the agreement between SourceFinder and a human rater ranged from 61% to 62%. This suggests that the estimated models have succeeded in capturing useful information about the characteristics of texts that affect test developers' ratings of source acceptability and that test developers may be able to use the system to find more high quality sources in less time. Since the process of locating acceptable source material is one of the most time-consuming parts of the item development process, this increase should translate directly into efficiency gains. In future work, we plan to further extend the system to include ratings appropriate for other types of verbal reasoning passages.

8. References

[1] Collins-Thompson, K. & Callan, J. "Predicting Reading Difficulty with Statistical Language Models," Journal of the American Society for Information Science and Technology, *56(13)*, 1448-1462, 2005.

[2] Petersen, S.E., & Ostendorf, M., "A Machine Learning Approach to Reading Level Assessment," University of Washington CSE Technical Report, 2006.

[3] Biber, D., Variation across Speech and Writing. Cambridge: Cambridge University Press, 1988.

[4] Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., Cortes, V., Csomay, E., & Urzua, A., "Representing Language Use in the University: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus," Educational Testing Service, Princeton, NJ TOEFL Monograph Series, 25, 2004.

[5] Salton, G. & McGill, M.J., Introduction to Modern Information Retrieval, McGraw Hill, 1983.

[6] Salton G. , Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. NY: Addison-Wesley Publishing, 1989.

[7] Burstein, J., The E-rater Scoring Engine. In M.D. Shermis and J.C. Burstein, Eds., Automated Essay Scoring: A Cross-Disciplinary Perspective. Mahwah, NJ: Lawrence Earlbaum Associates, 113-122, 2003.

[8] Lin, D., "Automatic retrieval and clustering of similar words," In Proceedings of the 35th Annual Meeting of the ACL, 898-901, 1998