Reading Difficulty in Adults with Intellectual Disabilities: Analysis with a Hierarchical Latent Trait Model

Martin Jansche
Google, Inc.
New York, NY, USA
jansche@acm.org

Lijun Feng
City University of New York
Graduate Center
New York, NY, USA
lijun7.feng@gmail.com

Matt Huenerfauth
City University of New York
Queens College & Graduate Center
New York, NY, USA
matt@cs.qc.cuny.edu

ABSTRACT

In prior work, adults with intellectual disabilities answered comprehension questions after reading texts. We apply a latent trait model to this data to infer the intrinsic difficulty of texts for the participant group. We then analyze the correlation between grade levels predicted by an automatic readability assessment tool and the inferred text difficulty.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Correlation and regression analysis; K.4.2 [**Computers and Society**]: Social Issues—*assistive technologies for persons with disabilities*

General Terms

Design, Experimentation, Measurement

Keywords

Assistive Technology, Intellectual Disabilities, Text Readability Assessment, Text Comprehension

1. INTRODUCTION

In recent work [4] we compared techniques for evaluating an envisioned text simplification system intended for adults with mild intellectual disabilities (ID). We then developed an automatic text readability assessment tool that predicts an appropriate grade level for a text [2]. This tool was developed and evaluated using a corpus of texts aimed at primary school students and annotated with grade levels. Adult readers with ID would also benefit from an automatic readability assessment tool, for reasons laid out in [1, 4]. Adapting, evaluating, and refining our assessment tool for this purpose requires an independent determination of the difficulty particular texts pose for adult readers with ID.

Here we apply a latent trait model to a subset of variables gathered in a previous reading experiment with adult participants with ID [4]. Relevant details of the experiment are described in Section 2; for more background, see [4]. We use this model to infer participants' ability levels and the intrinsic difficulties of particular texts they read and answered questions about. We then analyze the correlation between the inferred text difficulty and the grade levels predicted by our automatic readability assessment tool [2].

ASSETS'10, October 25–27, 2010, Orlando, Florida, USA.
ACM 978-1-60558-881-0/10/10.

2. EXPERIMENT AND DATA

The reading material for the experiment conducted in [4] consists of 11 original newswire articles and 11 corresponding simplified versions. The original articles were selected from local news to ensure familiarity. The simplified versions were manually adapted by a human expert specifically for adult readers with ID. Participants were asked to read the articles and answer 6 basic factual comprehension questions for each article. For each simplified article, the questions were identical to the corresponding original version.

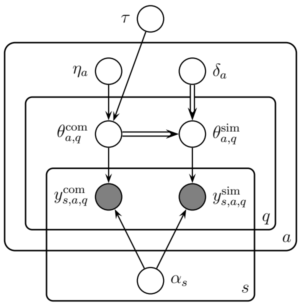
We recruited 20 adults with ID to participate in the experiment. Each participant was assigned 11 articles to read, some in their original and some in their simplified version. We made sure that no test participant saw both the original and simplified version of an article. The order of the articles and questions was randomized for each participant. The assignment of conditions – original vs. simplified version – was randomized with a margin constraint to ensure that all articles under both conditions would be presented to an equal number of participants.

Many more details of the experiment can be found in [4]. Here we concentrate on the participants' responses to the comprehension questions and what they tell us about the participants' abilities and the intrinsic difficulty of each article. Each observation recorded in the experiment consists of the participant number $s \in \{1, \ldots, 20\}$, the article topic $a \in \{1, \ldots, 11\}$, the version of the article $v \in \{\mathrm{com}, \mathrm{sim}\}$ (complex/original vs. simplified), the question number $q \in \{1, \ldots, 6\}$, and an indicator $y^v_{s,a,q} \in \{0, 1\}$ of whether the participant's response to the comprehension question was correct (1) or incorrect (0). The total number of $1320 = 20 \times 11 \times 6$ observations is a consequence of the design where each of the 20 participants read 11 texts and answered 6 questions per text. Due to time constraints during the experiment, only 1296 observations were collected.

3. MODEL AND COMPUTATION

The above presentation of the experiment in terms of stimuli and question responses that are either correct or incorrect immediately suggests an analysis based on an item response model or latent trait model. A direct application of the Rasch model to our data assumes a univariate latent trait which expresses both the abilities α of participants and the difficulties θ of question items (see e.g. §14.3 of [3]).

$$\Pr(y^v_{s,a,q} = 1) = \mathrm{logit}^{-1}(\alpha_s - \theta_{v,a,q})$$
$$\alpha_s \sim \mathrm{Normal}(\mu_\alpha, \sigma^2_\alpha) \qquad \theta_{v,a,q} \sim \mathrm{Normal}(\mu_\theta, \sigma^2_\theta)$$



$$\tau \sim \text{Gamma}(0.001, 0.001)$$

$$\eta_a \sim \text{N}(0, 10^{-6});\ \delta_a \sim \text{N}(0, 10^{-6})$$

$$\theta^{\text{com}}_{a,q} \sim \text{N}(\eta_a, \tau);\ \theta^{\text{sim}}_{a,q} = \theta^{\text{com}}_{a,q} - \delta_a$$

$$y^v_{s,a,q} \sim \text{Bern}(g^{-1}(\alpha_s - \theta^v_{a,q}))$$

$$\alpha_s \sim \text{N}(0, 1)$$

Figure 1: Hierarchical latent trait model

This model only captures some of the hierarchical structure inherent in the experimental design: each participant s is given a latent ability parameter α_s and each question item has a latent difficulty parameter. To the extent that a participant's ability exceeds an item's difficulty, the participant is more likely to answer the item correctly. More precisely, the inverse of the logit link function transforms the difference between ability and difficulty to a probability, where a difference of zero means the participant has equal chance of answering the question correctly or incorrectly.

We now enrich this basic model with an additional hierarchical structure to capture two additional aspects of the experimental design. First, items are no longer independent, but are grouped by article and condition. Second, our model will reflect the fact that the set of comprehension questions for each article was identical for the complex and simplified versions. We express this hierarchical structure in terms of additional latent variables in our model. Specifically, we assume:

- For each article a, a latent difficulty η_a. This can be thought of as the intrinsic difficulty of the original (complex) article.
- For each article a, a latent simplification amount δ_a. This expresses the reduction in difficulty when going from the original (complex) article to its simplified variant.
- For each article a and each associated question q, latent item difficulties $\theta^{\text{com}}_{a,q}$ and $\theta^{\text{sim}}_{a,q}$ for the complex and simplified versions, respectively, of the article.
- For each participant s, a latent ability α_s, as above in the Rasch model.

The full model then has the form shown in Figure 1. Here we follow the conventions of the BUGS language [7] and assume that normal distributions are parameterized in terms of mean and precision. To save space, we write N for a normal distribution and g for the logit link function. To ensure identifiability, we assume that the mean of the abilities α is known and fixed at zero.

The key property of our model lies in the structure it imposes on item-level difficulties. We assume that each original/complex article has an inherent difficulty η_a. The item-level difficulties $\theta^{\text{com}}_{a,q}$ for the original version of the article are drawn from a normal distribution with mean η_a. For the simplified version of the article, we ask the exact same questions, hence we assume that the item-level difficulty $\theta^{\text{sim}}_{a,q}$ of each question is reduced by the same article-level amount δ_a, representing the reduction in difficulty due to the simplification of the article. The observed responses are assumed to be generated by a standard Rasch model that combines participant abilities and item difficulties.

The model was formally specified in the BUGS language. Computations were carried out by Gibbs sampling using the JAGS software package [5], an open-source implementation very similar to classic BUGS. We used 3 parallel Markov chains that ran for 10,000 iterations each, which took about three minutes on a Linux workstation with a 3.16 GHz Intel Core2 CPU. The first 5,000 iterations in each chain were discarded, after checking for approximate convergence. We monitored all unobserved variables and noted that the potential scale reduction factor $\hat{R}$ was less than 1.03 in all cases, indicating approximate convergence (see e.g. §16 of [3]). The last 5,000 iterations in each chain were recorded and analyzed using the CODA [6] package for R. We checked the fit of the model by comparing the observed mean correct responses per article and per participant against the corresponding expected values under the posterior predictive distribution and found them to be uniformly close, revealing no obvious discrepancies between model and data.

4. RESULTS AND CONCLUSION

For our work on readability assessment, we were primarily interested in the quantities η_a (the difficulties of the 11 original articles), and $\eta_a - \delta_a$, which we take as the difficulties of the simplified articles. We used posterior means of each of these 22 quantities as point estimates for subsequent computations.

Adapting a readability assessment tool for adults with ID is complicated by the fact that there are no large-scale text corpora annotated with difficulty levels for this group of readers. Several small text corpora with grade-level annotations are available, and much larger amounts of data from educational testing could potentially be harnessed. We trained and evaluated our assessment tool [2] on texts annotated with primary school grade levels. It achieves a classification accuracy of 68%. We applied the tool to each of the 22 texts (original and simplified articles) and computed the correlation between predicted grade levels and text difficulty inferred by the model above. We found a correlation (Pearon's R) of 0.52. This compares favorably with expert ratings: when we asked 3 experts (two linguists and one psychology graduate student who has worked with people with ID) to rate the readability of the same texts on a 5-point scale, the correlations of their ratings with inferred text difficulty were 0.26, 0.14, and 0.03.

A hierarchical latent trait model is generally useful for inferring article-level difficulty from repeated observations based on multiple questions per article. This deserves to become as ubiquitous in research on adults with ID as it already is in educational testing and other settings.

5. REFERENCES

[1] L. Feng. Automatic readability assessment for people with intellectual disabilities. In *ASSETS 10*, 2008.

[2] L. Feng, M. Jansche, M. Huenerfauth, and N. Elhadad. A comparison of features for automatic readability assessment. In *COLING 23*, 2010.

[3] A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge, 2007.

[4] M. Huenerfauth, L. Feng, and N. Elhadad. Comparing evaluation techniques for text readability software for adults with intellectual disabilities. In *ASSETS 11*, 2009.

[5] M. Plummer. JAGS, version 2.1.0, May 2010.

[6] M. Plummer et al. CODA: Convergence diagnosis and output analysis for MCMC. *R News*, 6(1):7–11, 2006.

[7] A. Thomas. The BUGS language. *R News*, 6(1):17–21, 2006.