Finding your "inner-annotator": An experiment in annotator independence for rating discourse coherence quality in essays

Jill Burstein
Educational Testing Service
666 Rosedale Road
Princeton, NJ 08541

Swapna Somasundaran
Educational Testing Service
666 Rosedale Road
Princeton, NJ 08541

Martin Chodorow
Hunter College, CUNY
695 Park Avenue
New York, NY

jburstein@ets.org ssomasundaran@ets.org martin.chodorow@hunter.cuny.edu

Abstract

An experimental annotation method is described, showing promise for a subjective labeling task – discourse coherence quality of essays. Annotators developed personal protocols, reducing front-end resources: protocol development and annotator training. Substantial inter-annotator agreement was achieved for a 4-point scale. Correlational analyses revealed how unique linguistic phenomena were considered in annotation. Systems trained with the annotator data demonstrated utility of the data.

1 Introduction[1]

Systems designed to evaluate discourse coherence quality often use supervised methods, relying on human annotation that requires significant front-end resources (time and cost) for protocol development and annotator training (Burstein et al., 2013). Crowd-sourcing (e.g., Amazon Mechanical Turk) has been used to collect annotation judgments more efficiently than traditional means for tasks requiring little domain expertise (Beigman Klebanov et al., 2013; Louis & Nenkova, 2013). However, proprietary data (test-taker essays) may preclude crowd-sourcing use. In the U.S., the need for automated writing evaluation systems to score proprietary test-taker data is likely to increase when Common Core[2] assessments are administered to school-age students beginning in 2015 (Shermis, in press), increasing the need for data annotation. This paper describes an experimental method for capturing discourse coherence quality judgments for test-taker essays. Annotators developed personal protocols reflecting their intuitions about essay coherence, thus reducing standard front-end resources. The paper presents related work (Section 2), the experimental annotation (Section 3), system evaluations (Section 4), and conclusions (Section 5).

2 Related Work

Even after extensive training, subjective tasks may yield low inter-annotator agreement (Burstein & Wolska, 2003; Reidsma & op den Akker, 2008; Burstein et al., 2013). Front-end annotation activities may require significant resources (protocol development and annotator training) (Miltsakaki and Kukich, 2000; Higgins, et al., 2004; Wang et al., 2012; Burstein et al., 2013). Burstein et al (2013) reviewed coherence features as discussed in cognitive psychology (Graesser et al., 2004), reading research (Van den Broek, 2012), and computational linguistics, and concluded that evaluating text coherence is *highly personal* , relying on a *variety of features*, including adherence to standard writing conventions (e.g., grammar), and patterns of rhetorical structure and vocabulary usage. They describe an annotation protocol that uses a 3-point coherence quality scale (3 (*high*), 2 (*somewhat*,) and 1 (*low*)) applied by 2 annotators to label 1,500 test-taker essays from 6 task types (Table 1). Protocol development took several weeks, and offered extensive descriptions of the 3-point scale, including illustrative test-taker responses; rigorous annotator training was also conducted. Burstein et al, 2013 collapsing the 3-point scale to a 2-point scale (i.e., *high* (3), *low* (1,2)). Results for a *binary* discourse coherence quality system (*high* and *low* coherence) for essays achieved only *borderline modest*

[2] See http://www.corestandards.org/.

LAW VIII - The 8th Linguistic Annotation Workshop, pages 48–53,
Dublin, Ireland, August 23-24 2014.

Essay-Writing Item Type	Test-Taker Population
1. K-12 expository	Students[3], ages 11-16
2. Expository	NNES-Univ
3. Source-based, integrated (reading and listening)	NNES-Univ
4. Expository	Graduate school applicants
5. Critical argument	Graduate school applicants
6. Professional licensing, content/expository	Certification for a business-related profession

Table 1. Six item types & populations in the experimental annotation task. NNES-Univ = non-native English speakers, university applicants

performance (κ=0.41)[4]. Outcomes reported in Burstein et al are consistent with discussions that text coherence is a complex and individual process (Graesser et al, 2004; Van den Broek, 2012), motivating our experimental method. In contrast to training annotators to follow an annotation scheme pre-determined by others, annotators devised their own scoring protocols, capturing their independent impressions – *finding their "inner-annotator*." The practical outcomes of success of the method would be reduced front-end resources in terms of time required to (a) develop the annotation protocol and (b) train annotators. As a practical end-goal, another success criterion would be to achieve inter-annotator agreement such that classifiers could be trained, yielding *substantial* annotator-system agreement.

3 Experimental Annotation Study

Annotation scoring protocols from 2 annotators for coherence quality are evaluated and described.

3.1 Human Annotators

Two high school English teachers (employed by a company specializing in annotation) performed the annotation. Annotators never met each other, did not know about each other's activities, and only communicated about the annotation with a facilitator from the company.

3.2 Data

A random sample of 250 essays for 6 different item types (*n*=1500) and test-taker populations (Table 1) was selected. The sample was selected across 20 different prompts (test questions) for each item type in order to ensure topic generalizability in the resulting systems. Forty essays were randomly selected for a small pilot study; the remaining data (1460 essays) were used for the full annotation study. For the full study, 20% of the essays (*n*=292) had been randomly selected for double annotation to measure inter-annotator agreement; the remaining 1168 essays were evenly divided, and each annotator labeled half (*n*=584 per annotator). Each annotator labeled a total of 876 essays across the 6 task types.

3.3 Experimental Method Description

A one-week pilot study was conducted. To provide some initial grounding, annotators received a 1-page task description that offered a high-level explanation of "coherence" describing the end-points of a potential protocol. (This description was written in about an hour.) It indicated that high coherence is associated with an essay that can be *easily understood*, and low coherence is associated with an *incomprehensible* essay. Each annotator developed her own protocol: for each score point she wrote descriptive text illustrating a set of defining characteristics for each score point of coherence quality (e.g., "*The writer's point is difficult to understand*."). Annotator 1 (A1) developed a 4-point scale;

[3] Note that this task type was administered in an instructional setting; all other tasks were completed in high-stakes assessment settings.

[4] Kappa was not reported in the paper, but was accessed through personal communication.

Feature Type	**A1 (*r*)**	**A2 (*r*)**
Grammar errors (e.g., subject verb agreement)	0.42	0.35
Word usage errors (e.g., determiner errors)	0.46	0.44
Mechanics errors (e.g., spelling, punctuation)	0.58	0.52
EGT -- best 3 features (out of 112 features): F1, F2, F3	F1. -0.30 F2. -0.28 F3. 0.27	F1. -0.14 F2. -0.15 F3. 0.11
RST features-- best 3 features (out of 100 features): F1, F2, F3	F1. -0.27 F2. 0.15 F3. 0.19	F1. -0.19 F2. 0.08 F3. 0.06
LDSP	0.19	0.06

Table 2. Pearson *r* between annotator discourse coherence scores and features. All correlations are significant at $p < .0001$, except for A2's long-distance sentence-pair similarity at $p < .05$.

Annotator 2 (A2) developed a 5-point scale. Because the two scales were different, κ could not be used to measure agreement, so a Spearman rank-order correlation (r_S) was used, yielding a promising value (r_S=0.82). Annotator protocols were completed at the end of the pilot study.

A full experiment was conducted. Each annotator used her protocol to assign a coherence quality score to each essay. Annotators assigned a score and wrote brief comments as explanation (drawing from the protocol). Comments provided a score supplement that could be used to support analyses beyond quantitative measures (Reidsma & Carletta, 2008). The data were annotated in 12 batches (by task) composed of 75 essays (50 unique; 25 for double annotation). A Spearman rank-order correlation was computed on the double-scored essays for completed batches. If the correlation fell below 0.70 (which was infrequent), one of the authors reviewed the annotator scores and comments to look for inconsistencies. Agreement was re-computed when annotator revisions were completed to ensure inter-rater agreement of 0.70. Annotations were completed over approximately 4 weeks to accommodate annotator schedules. While a time log was not strictly maintained, we estimate the total time for communication to resolve inconsistency issues was about 4-6 hours. One author communicated *score-comment inconsistencies* (e.g., high score with critical comments) to the company's facilitator (through a brief e-mail); the facilitator then relayed the inconsistency information to the annotator(s). The author's data review and communication e-mail took no longer than 45 minutes for the few rounds where agreement fell below 0.70. Communication between the facilitator and the annotator(s) involved a brief discussion, essentially reviewing the points made in the e-mail.

3.4 Results: Inter-annotator agreement

Using the Spearman rank-order correlation, inter-rater agreement on the double-annotated data was r_S=0.71. In order to calculate Kappa statistic, A2's 5-point scale assignments were then mapped to a 4-point scale by collapsing the two lowest categories (1,2) into one (1), since there were very few cases of 1's; this is consistent with low frequencies of very low-scoring essays. Using quadratic weighted kappa (QWK), post-mapping indicated *substantial* agreement between the two annotators (κ=0.61).

3.5 Correlational Analysis: Which Linguistic Features Did Annotators Consider?

A1 and A2 wrote brief comments explaining their coherence scores. Comments were shorthand notation drawn from their protocols (e.g., *There are significant grammatical errors...thoughts do not connect.*). Both annotators included descriptions such as "*word patterns*," "*logical sequencing*," and "*clarity of ideas*"; however, A2 appeared to have more comments related to grammar and spelling. Burstein et al., (2013) describe the following features in their binary classification system: (1) grammar, word usage, and mechanics errors (GUM), (2) rhetorical parse tree features (Marcu, 2000) (RST), (3) entity-grid transition probabilities to capture local "topic distribution" (Barzilay & Lapata, 2008) (EGT), and (4) a long-distance sentence pair similarity measure using latent semantic analysis (Foltz, 1998) to capture "long distance, topical distribution. (LDSP). Annotated data from this study were processed with the Burstein et al (2013) system to extract the features above in (1) – (4). To quantify the observed

differences in the annotators' comments and potential effects for system score assignment (Section 4), we computed Pearson (*r*) correlations between the system features (on our annotated data set), and the discourse coherence scores of A1 and A2 (using the 4-point scale mapping for A2). There are 112 entity-transition probability features and 100 Rhetorical Structure Theory (RST) features. In Table 2, the correlations of the three best predictors from the EGT and RST sets, and the GUM features and the LDSP feature are shown. Correlations in Table 2 are significantly correlated between the feature sets and annotator coherence scores. However, we observed that the EGT, RST, and LDSP feature correlation values for A2 are notably smaller than A1's. This suggests that A2 may have had a strong reliance on GUM features, or that the system feature set did not capture all linguistic phenomena that A2 considered.

4 System Evaluation[5]

To evaluate the utility of the annotated data, two evaluations were conducted: one built classifiers with all system features (Sys_All), and a second with the GUM features (Sys_GUM). Using 10-fold cross-validation with a gradient boosting regression learner, four classifiers were trained to predict coherence quality ratings on a 4-point scale, using the respective annotator data sets: A1 and A2 Sys_All, and A1 and A2 Sys_GUM systems.

4.1 Results

Sys-All trained with A1 data consistently outperformed Sys-All trained with A2 data. Results are reported for averages across the 10-folds, and showed *substantial* system-human agreement for A1 (κ = 0.68) and *modest* system-human agreement for A2 (κ = 0.55). When Sys_GUM was trained with A1 data, system-human agreement dropped to a *modest* range (κ = 0.60); when Sys_GUM was trained with A2 data, however, human agreement was essentially unchanged, staying in the *modest* agreement range (κ = 0.50). Consistent with the correlational analysis, this finding suggests that A2 has strong reliance on GUM features, or the system may have been less successful in capturing A2 features beyond GUM.

5 Discussion and Conclusions

Our experimental annotation method significantly reduced front-end resources for protocol development and annotator training. Analyses reflect one genre: essays from standardized assessments. Minimal time was required from the authors or the facilitator (about two hours) for protocol development; the annotators developed personal protocols over a week during the pilot; in Burstein et al (2013), this process was report to take about one month. Approximately 4-6 hours of additional discussion from one author and the facilitator was required during the task; Burstein et al (2013) required two researchers and two annotators participated in several 4-hour training sessions, totaling about 64-80 hours of person-time across the 4 participants (personal communication). In addition to its efficiency, the experimental method was *successful* per criteria in Section 2. The method captures annotators' subjective judgments about coherence quality, yielding *substantial* inter-annotator agreement (κ=0.61) across a 4-point scale. Second, classifiers trained with annotator data showed that the systems showed *substantial* and *modest* agreement (A1 and A2, respectively) – demonstrating annotation utility, especially for A1. Correlational analyses were used to analyze effects of features that annotators may have considered in making their decisions. Comment patterns and results from the correlation analysis suggested that A2's decisions were either based on narrower considerations (GUM errors), or not captured by our feature set.

The experimental task facilitated the successful collection of subjective coherence judgments with substantial inter-annotator agreement on test-taker essays. Consistent with conclusions from Reidsma & Carletta (2008), outcomes show that quantitative measures of inter-annotator agreement should not be used exclusively. Descriptive comments were useful for monitoring *during* annotation, interpreting annotator considerations and system evaluations *during* and *after* annotation, and informing system development. In the future, we would explore strategies to evaluate intra-annotator reliability (Beigman-Klebanov, Beigman, & Diermeier, 2008) which may have contributed to lower system performance with A2 data.

[5] Many thanks to Binod Gywali for engineering support.

References

Beata Beigman-Klebanov, Nitin Madnani,, and Jill Burstein. 2013. Using Pivot-Based Paraphrasing and Sentiment Profiles to Improve a Subjectivity Lexicon for Essay Data, *Transactions of the Association for Computational Linguistic*s, Vol.1: 99-110.

Beata Beigman-Klebanov, Eyal Beigman, and Daniel Diermeier. 2008. Analyzing Disagreements. In Proceedings of the workshop on Human Judgments in Computational Linguistics, Manchester: 2-7.

Jill Burstein, Joel Tetreault and Martin Chodorow. 2013. Holistic Annotation of Discourse Coherence Quality in Noisy Essay Writing. In the *Special issue of Dialogue and Discourse on: Beyond semantics: the challenges of annotating pragmatic and discourse phenomena* Eds. S. Dipper, H. Zinsmeister, and B. Webber. Discourse & Dialogue 42, 34-52.

Jill Burstein and Magdalena Wolska..2003. Toward Evaluation of Writing Style: Overly Repetitious Word Use. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*. Budapest, Hungary.

Jacob Cohen. 1960. "A coefficient of agreement for nominal scales". Educational and Psychological Measurement 20 1: 37–46.

Joseph Fleiss and Jacob Cohen 1973. "The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability" in *Educational and Psychological Measurement*, Vol. 33:613–619.

Peter Foltz, Walter Kintsch, & Thomas Landuaer. 1998. Textual coherence using latent semantic analysis. *Discourse Processes*, 252&3: 285–307.

Arthur Graesser, Danielle McNamara, Max Louwerse. and Zhiqiang Cai, Z. 2004. Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2), 193-202.

Derrick Higgins, Jill Burstein, Daniel Marcu &. Claudia Gentile. 2004. Evaluating Multiple Aspects of Coherence in Student Essays. In *Proceedings of 4th Annual Meeting of the Human Language Technology and North American Association for Computation Linguistics*:185–192, Boston, MA

J. Richard Landis,. & G. Koch. 1977. "The measurement of observer agreement for categorical data". *Biometrics* **33** 1: 159–174.

Annie Louis and Ani Nenkova. 2013. A Text Quality Corpus for Science Journalism. In the *Special Issue of Dialogue and Discourse on: Beyond semantics: the challenges of annotating pragmatic and discourse phenomena* Eds. S. Dipper, H. Zinsmeister, and B. Webber, 42: 87-117.

Eleni Miltsakaki and Karen Kukich. 2000. Automated evaluation of coherence in student essays. In *Proceedings of the Language Resources and Evaluation Conference*, Athens, Greece.

Daniel Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization.* Cambridge, MA: The MIT Press.

Dennis Reidsma, and Rieks op den Akker. 2008. Exploiting `Subjective' Annotations. In *Proceedings of the Workshop on Human Judgments in Computational Linguistics*, Coling 2008, 23 August 2008, Manchester, UK.

Dennis Reidsma, and Jean Carletta. 2008. Reliability measurements without limits. *Computational Linguistics*, 343: 319-336.

Mark Shermis. to appear. State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*.

Y. Wang, M. Harrington, and P. White. 2012. Detecting Breakdowns in Local Coherence in the Writing of Chinese English Speakers. The *Journal of Computer Assisted Learning*. 28: 396–410.

Paul Van den Broek. 2012. Individual and developmental differences in reading comprehension: Assessing cognitive processes and outcomes. In: Sabatini, J.P., Albro, E.R., O'Reilly, T. (Eds.), *Measuring up: Advances in how we assess reading ability.*, pp. 39-58. Lanham: Rowman & Littlefield Education.