Using Entity-Based Features to Model Coherence in Student Essays

Jill Burstein
Educational Testing Service
Princeton, NJ 08541
jburstein@ets.org

Joel Tetreault
Educational Testing Service
Princeton, NJ 08541
jtetreault@ets.org

Slava Andreyev
Educational Testing Service
Princeton, NJ 08541
sandreyev@ets.org

Abstract

We show how the Barzilay and Lapata entity-based coherence algorithm (2008) can be applied to a new, noisy data domain – *student essays*. We demonstrate that by combining Barzilay and Lapata's entity-based features with novel features related to grammar errors and word usage, one can greatly improve the performance of automated coherence prediction for student essays for different populations.

1 Introduction

There is a small body of work that has investigated using NLP for the problem of identifying coherence in student essays. For example, Foltz, Kintsch & Landauer (1998), and Higgins, Burstein, Marcu & Gentile (2004) have developed systems that examine coherence in student writing. Foltz, *et al*. (1998) systems measure lexical relatedness between text segments by using vector-based similarity between adjacent sentences; Higgins *et al*'s (2004) system computes similarity across text segments. Foltz *et al*.'s (1998) approach is in line with the earlier TextTiling method that identifies subtopic structure in text (Hearst, 1997). Miltsakaki and Kukich (2000) addressed essay coherence using Centering Theory (Grosz, Joshi & Weinstein, 1995). More recently, Barzilay and Lapata's (2008) approach (henceforth, BL08) used an entity-based representation to evaluate coherence. In BL08, *entities* (nouns and pronouns) are represented by their sentence roles in a text. The algorithm keeps track of the distribution of entity transitions between adjacent sentences, and computes a value for all transition types based on their proportion of occurrence in a text. BL08 apply their algorithm to three tasks, using well-formed newspaper corpora: text ordering, summary coherence evaluation, and readability assessment. For each task, their system outperforms a Latent Semantic Analysis baseline. In addition, best performance on each task is achieved using different system and feature configurations. Pitler & Nenkova (2008) applied BL08 to detect text coherence in well-formed texts.

Coherence quality is typically present in scoring criteria for evaluating a student's essay. This paper focuses on the development of models to predict *low*-and *high-coherence ratings* for essays. Student essay data, unlike newspaper text, is typically noisy, especially when students are non-native English speakers (NNES). Here, we evaluate how BL08 algorithm features can be used to model coherence in a new, noisy data domain -- *student essays*. We found that coherence can be best modeled by combining BL08 entity-based features with novel writing quality features. Further, our use of data sets from three different test-taker populations also shows that coherence models will differ across populations. Different populations might use language differently which could affect how coherence is presented. We expect to incorporate coherence ratings into e-rater®, ETS's automated essay scoring system (Attali & Burstein, 2006).

2 Corpus and Annotation

We collected approximately 800 essays (in total) across three data sets[1]: 1) adult, NNES test essays (TOEFL); 2) adult, native and NNES test essays; (GRE) 3) U.S. middle- and high-school native and NNES student essay submissions to *Criterion*®, ETS's instructional writing application.

Two annotators were trained to rate *coherence quality* based on how easily they could read an essay without stumbling on a *coherence barrier* (i.e., a confusing sentence(s)). Annotators rated

[1] TOEFL® is the Test of English as a Foreign Language, and GRE® is the Graduate Record Admissions Test.

essays on a 3-point scale: 1) *low coherence*, 2) *somewhat coherent*, and 3) *high coherence*. They were instructed to ignore grammar and spelling errors, unless they affected essay comprehension.

During training, *Kappa* agreement statistics indicated that annotators had difficulty agreeing on the middle, *somewhat coherent* category. The annotation scale was therefore collapsed into a 2-point scale: *somewhat coherent* and *high coherence* categories were collapsed into the *high coherence* class (H), and *low-coherence* (L) remained unchanged. Two annotators labeled an overlapping set of about 100 essays to calculate inter-rater agreement; weighted Kappa was 0.677.

3 System

3.1 BL08 Algorithm

We implemented BL08's entity-based algorithm to build and evaluate coherence models for the essay data. In short, the algorithm generates a vector of *entity transition probabilities* for documents (essays, here). Vectors are used to build coherence models. The first step in the algorithm is to construct an entity grid in which all entities (*nouns and pronouns*) are represented by their roles (i.e., Subject **(S)**, Object **(O),** Other **(X)).** Entity roles are then used to generate *entity transitions* – the role transitions across adjacent sentences (e.g., Subject-to-Object, Object-to-Object). *Entity transition probabilities* are the proportions of different entity transition types within a text. The probability values are used then used as features to build a coherence model.

Entity roles can be represented in the following ways. In this study, consistent with BL08, different combinations are applied and reported (see Tables 2-4). Entities can be represented in grids with specified roles **(Syntax+)** (**S,O,X**). Alternatively, roles can be reduced to show *only* the presence and absence of an entity **(Syntax-)** (i.e., Entity Present **(P)** or Not **(N).** Co-referential entities can be resolved **(Coreference+)** or not **(Coreference-).** Finally, the **Salience** option reflects the frequency with which an entity appears in the discourse: if the entity is mentioned two or more times, it is salient **(Salient+),** otherwise, not **(Salient-)**.

Consistent with BL08, we systematically completed runs using various configurations of entity representations (see Section 4).

Given the combination, the entity transition probabilities were computed for all labeled essays in each data set. We used *n-fold* cross-validation for evaluation. Feature vectors were input to C5.0, a decision-tree machine learning application.

3.2 Additional Features

In BL08, augmenting the core coherence features with additional features improved the power of the algorithm. We extended the feature set with writing quality features (Table 1). **GUMS** features describe the technical quality of the essay. The motivation for type/token features **(*_TT)** is to measure word variety. For example, a high probability for a "Subject-to-Subject" transition indicates that the writer is repeating an entity in Subject position across adjacent sentences. However, this does not take into account whether the same word is repeated or a variety of words are used. The **{S,O,X,SOX}_TT** (type/token) features *uncover* the actual words collapsed into the entity transition probabilities. **Shell nouns** (Atkas & Cortes, 2008), common in essay writing, might also affect coherence.

NNES essays can contain many spelling errors. We evaluated the impact of a context-sensitive spell checker **(SPCR+),** as spelling variation will affect the transition probabilities in the entity grid.

Finally, we experimented with a *majority vote method* that combined the best performing feature combinations.

4 Evaluation

For all experiments, we used a series of *n-fold cross-validation* runs with C5.0 to evaluate performance for numerous feature configurations. In Tables 2, 3 and 4, we report: baselines, results on our data with BL08's best system configuration from the summary coherence evaluation task (closest to our task), and our best systems. In the Tables, "best systems" combined feature sets and outperformed baselines. Rows in **bold** indicate final independent best systems that contribute to best performance in the *majority vote* method. Agreement is reported as Weighted Kappa (**WK**), Precision (**P**), Recall (**R**) and F-measure (**F**).

Baselines. We implemented three non-trivial baseline systems. **E-rater** indicates use of the full

feature set from e-rater. The **GUMS** (GUMS+) feature baseline, uses the Grammar (G+), Usage

Feature Descriptor	Feature Description
GUMS	Grammar, usage, and mechanics errors, and style features from an AES system
S_TT **O_TT** **X_TT** **SOX_TT**[2] **P_TT**	Type/token ratios for actual words recovered from the entity grid, using the entity *roles*.
S_TT_Shellnouns **O_TT_Shellnouns** **X_TT_Shellnouns**	Type/token ratio of non-topic content, *shell nouns* (e.g., *approach, aspect, challenge*)

Table 1: New feature category description

(U+), Mechanics (M+), and Style (ST+) flags (subset of e-rater features) to evaluate a coherence model. The third baseline represents the best run using **type/token features** ({S,O,X,SOX}_TT), and **{S,O,X}_TT_Shellnouns** feature sets (Table 1). The baseline majority voting system includes e-rater, GUMS, and the best performing type/token baseline (see Tables 2-4).

Extended System. We combined our writing quality features with the core BL08 feature set. The combination improved performance over the three baselines, and over the best performing BL08 feature. Type/token features added to BL08 entity transitions probabilities improved performance of all single systems. This supports the need to *recover* actual word use. In Table 2, for TOEFL data, spell correction improved performance with the Mechanics error feature (where Spelling is evaluated). *This would suggest that annotators were trying to ignore spelling errors when labeling coherence.* In Table 3, for GRE data, spell correction improved performance with the Grammar error feature. *Spell correction did change grammar errors detected: annotators may have self-corrected for grammar.* Finally, the *majority vote* outperformed all systems. In Tables 3 and 4, Kappa was comparable to human agreement (K=0.677).

5 Conclusions and Future Work

We have evaluated how the BL08 algorithm features can be used to model coherence for student essays across three different populations. We found that the best coherence models for essays are built by combining BL08 entity-based features with writing quality features. BL08's outcomes showed that optimal performance was obtained by using different feature sets for different tasks. Our task was most similar to BL08's summary coherence task, but we used noisy essay data. The difference in the data types might also explain the need for our systems to include additional writing quality features.

Our *majority vote* method outperformed three baselines (and a baseline majority vote). For two of the populations, Weighted Kappa between system and human agreement was comparable. These results show promise toward development of an entity-based method that produces reliable coherence ratings for noisy essay data. We plan to evaluate this method on additional data sets, and in the context of automated essay scoring.

[2] Indicates an aggregate feature that computes the type/token ratio for entities that appear in any of S,O,X role.

References

Aktas, R. N., & Cortes, V. (2008). Shell nouns as cohesive devices in published and ESL student writing. *Journal of English for Academic Purposes*, 7(1), 3–14.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with *e-rater* v.2.0 . *Journal of Technology, Learning, and Assessment, 4*(3).

Barzilay, R. and Lapata, M. (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1), 1-34.

Foltz, P., Kintsch, W., and Landauer, T. K. (1998). The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes*, 25(2&3):285–307.

Higgins, D., Burstein, J., Marcu, D., & Gentile, C. (2004). Evaluating multiple aspects of coherence in student essays . *In Proceedings of HLT-NAACL 2004*, Boston, MA.

Grosz, B., Joshi, A., and Weinstein, S. 1995, Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2): 203-226.

Hearst, M. A. (1997). TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–6

Miltsakaki, E. and Kukich, K. (2000). Automated evaluation of coherence in student essays. In *Proceedings of LREC 2000*, Athens, Greece

Pitler, E.,and Nenkova, A (2008). Revisiting Readability: A Unified Framework for Predicting

Text Quality. *In Proceedings of EMNLP 2008*, Honolulu, Hawaii.

		L (n=64)			H (n=196)			L+H (n=260)		
BASELINES: NO BL08 FEATURES	**WK**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
(a) E-rater	0.472	56	69	62	89	82	86	79	79	79
(b) GUMS	0.455	55	66	60	88	83	85	79	79	79
(c) SOX_TT[3]	0.484	66	55	60	86	91	88	82	82	82
SYSTEMS: Includes BL08 FEATURES										
Coreference-Syntax+Salient+ (B&L08 summary task configuration)	0.253	49	34	40	81	88	84	75	75	75
(d) Coreference-Syntax-Salient-SPCR+M+	**0.472**	**76**	**45**	**57**	**84**	**95**	**90**	**83**	**83**	**83**
(e) Coreference+Syntax+Salient-GUMS+	**0.590**	**68**	**70**	**69**	**90**	**89**	**90**	**85**	**85**	**85**
(f) Coreference+Syntax+Salient-GUMS+O_TT_Shellnouns+	**0.595**	**68**	**72**	**70**	**91**	**89**	**90**	**85**	**85**	**85**
Baseline Majority vote: (a),(b), (c)	0.450	55	64	59	88	83	85	79	79	79
Majority vote: (d), (e), (f)	**0.598**	**69**	**70**	**70**	**90**	**90**	**90**	**85**	**85**	**85**

Table 2: Non-native English Speaker Test-taker Data (TOEFL): Annotator/System Agreement

		L (n=48)			H (n=210)			L+H (n=258)		
BASELINES: NO BL08 FEATURES	**WK**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
(a) E-rater	0.383	79	31	45	86	98	92	86	86	86
(b) GUMS	0.316	68	27	39	85	97	91	84	84	84
(c) e-rater+SOX_TT[4]	0.359	78	29	42	86	98	92	85	85	85
SYSTEMS: INCLUDES BL08 FEATURES										
Coreference-Syntax+Salient+ (BL08 summary task configuration)	0.120	35	17	23	83	93	88	79	79	79
(d) Coreference+Syntax+Salient-SPCR+G+	**0.547**	**1.0**	**43**	**60**	**89**	**1.0**	**94**	**90**	**90**	**90**
(e) Coreference+Syntax-Salient-P_TT+	**0.462**	**70**	**44**	**54**	**88**	**96**	**92**	**86**	**86**	**86**
(f) Coreference+Syntax+Salient+GUMS+ SOX_TT+	**0.580**	**71**	**60**	**65**	**91**	**94**	**93**	**88**	**88**	**88**
Baseline Majority vote: (a),(b), (c)	0.383	79	31	45	86	98	92	86	86	86
Majority vote: (d), (e), (f)	**0.610**	**1.0**	**49**	**66**	**90**	**1.0**	**95**	**91**	**91**	**91**

Table 3: Native and Non-Native English Speaker Test-taker Data (GRE): Annotator/System Agreement

		L (n=37)			H (n=226)			L+H (n=263)		
BASELINES: NO BL08 FEATURES	**WK**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
(a) E-rater	0.315	39	46	42	91	88	89	82	82	82
(b) GUMS	0.350	47	41	43	90	92	91	85	85	85
(c) SOX_TT	0.263	78	19	30	88	99	93	88	88	88
SYSTEMS: INCLUDES BL08 FEATURES										
(d) Coreference-Syntax+Salient+ (BL08 summary task configuration)	**0.383**	**79**	**30**	**43**	**90**	**99**	**94**	**89**	**89**	**89**
(e) Coreference-Syntax-Salient-SPCR+	**0.424**	**67**	**38**	**48**	**90**	**97**	**94**	**89**	**89**	**89**
(f) Coreference+Syntax+Salient+S_TT+	**0.439**	**65**	**41**	**50**	**91**	**96**	**94**	**89**	**89**	**89**
Baseline Majority vote: (a),(b), (c)	0.324	43	41	42	90	91	91	84	84	84
Majority vote: (d), (e), (f)	**0.471**	**82**	**38**	**52**	**91**	**99**	**94**	**90**	**90**	**90**

Table 4: *Criterion* Essay Data*:* Annotator/System Agreement

[3] Type/token ratios from all roles using a Coreference+Syntax+Salient+ grid.
[4] Type/token ratios from all roles using Coreference+Syntax+Salient- grid.