Topicality-Based Indices for Essay Scoring

Beata Beigman Klebanov, Michael Flor, and Binod Gyawali
Educational Testing Service
660 Rosedale Road
Princeton, NJ 08541
bbeigmanklebanov,mflor,bgyawali@ets.org

Abstract

In this paper, we address the problem of quantifying the overall extent to which a test-taker's essay deals with the topic it is assigned (prompt). We experiment with a number of models for word topicality, and a number of approaches for aggregating word-level indices into text-level ones. All models are evaluated for their ability to predict the holistic quality of essays. We show that the best text-topicality model provides a significant improvement in a state-of-art essay scoring system. We also show that the findings of the relative merits of different models generalize well across three different datasets.

1 Introduction

The instruction to "stay on topic" oft given to developing writers seems intuitively unproblematic, yet the question of the best way to measure this property of a text is far from settled, and little is known about the interaction of topicality and other properties of text, such as length. We develop text topicality indices and evaluate them in the context of automated scoring of essays. Specifically, we investigate the relationship between the extent to which the essay engages the topic provided in the essay question (prompt) and the quality of the essay as quantified by a human-provided holistic score.

In the existing literature, topicality has been addressed as a control flag to identify off-topic essays or spoken responses (Yoon and Xie, 2014; Louis and Higgins, 2010; Higgins et al., 2006) or as an element in the overall coherence of the essay (Somasundaran et al., 2014; Higgins et al., 2004; Foltz et al., 1998). Persing and Ng (2014) annotated essays for prompt-adherence, and found that achieving inter-rater reliability was very challenging, reporting Pearson r = 0.243 between two raters. We address the relationship between a continuous topicality score and the holistic quality of an essay.

Generally, one can think of the topicality of a given word w on a given topic T as the extent to which w occurs more often in texts addressing T than in otherwise comparable texts addressing a different topic. We consider three models of word topicality from the literature: the significance-test approach as in topic signatures (Lin and Hovy, 2000), the score-product approach as described in the essay scoring literature (Higgins et al., 2006), and a simple cutoff-based approach relying on difference in probabilities.

Given a definition of word topicality, the question arises how to quantify the topicality of the whole text. Specifically, is topicality a property of the vocabulary of a text (of word types) or a property of both the vocabulary and the unfolding discourse (of word tokens)? Thus, do the sentences "I hate restaurants, abhor restaurants, loath restaurants, and love restaurants" and "I hate restaurants, abhor waiters, loath menus, and love food" address the topic of restaurants to the same extent (this would be the prediction of the token-based model), or does the latter sentence address the topic to a greater extent than the former (this would be the prediction of the type-based model)?[1] The second sentence seems to en-

[1] Assuming the 4 verbs in the example are off-topic and the nouns are on-topic of restaurants, the token-based model

Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, pages 63–72,
San Diego, California, June 16, 2016. ©2016 Association for Computational Linguistics

gage more with the topic because it attends to more aspects (or details) of the topic.

In this paper, we implement type-based and token-based approaches to text topicality, using a number of different models for word topicality. All models are evaluated for their ability to predict the holistic quality of an essay.

The contributions of this paper are as follows. First, assuming a number of common definitions of word topicality and an application of predicting holistic quality of essays, we show that text-level topicality is most effectively modeled (a) as a property of word types rather than tokens in the text; (b) taking essay length into account. Second, we show that when word topicality is defined using a simple cutoff-based measure and text-topicality is modeled as in (a),(b) above, we obtain a predictor of essay score that yields a statistically significant improvement in a state-of-art essay scoring system. Third, we show that the characteristics of the best topicality model and its effectiveness in improving essay scoring generalize across different kinds of essays.

2 Data

We experiment with three datasets. Two are datasets of essays responding to two different essay tasks written for a large-scale college-level examination in the United States. These essays are scored by professional raters on a 6-point scale. These sets contain tens of thousands of essays responding to dozens different prompt questions (82,500 essays, 76 prompts for each dataset). Their sheer sizes and the variety of topics (prompts) allow for a thorough evaluation of the proposed measures. However, the proprietary nature of these data does not allow for easy replication of the results, or benchmarking; we therefore use a third, publicly available dataset containing 12,100 essays written for the TOEFL test by non-native speakers of English seeking college entrance in the United States, as well as for other purposes. The dataset was originally built for the task of native language identification (Blanchard et al., 2013; Tetreault et al., 2012); however, the distribution provides coarse-grained holistic scores as well (3-point scale). We describe each of the datasets in detail below. Table 1 shows the sizes of the partitions of the datasets into Dev (used for building topicality models), Train (used for selecting the best topicality model and for training the essay scoring system); Test (used for a blind test of the essay scoring system). Table 2 shows score distributions and mean essay length on Train data.

Part.	Set 1	Set 2	TOEFL
Dev	76 × 500	76 × 500	8 × 500
Train	51 × 500	51 × 500	8 × 760 (Av)
Test	76 × 250	76 × 250	8 × 253 (Av)

Table 1: Sizes of the data partitions for each dataset. In the $N \times M$ notation, N = # prompts, M = # essays per prompt. In TOEFL train and test sets, we show average numbers of essays per prompt.

Score	Set 1	Set 2	Score	TOEFL
1	0.015	0.015	low	0.108
2	0.154	0.123	med	0.546
3	0.384	0.412	high	0.346
4	0.327	0.342	—	
5	0.104	0.096	—	
6	0.016	0.012	—	
Av. Len.	395	405	Av. Len.	317
(Std.)	(129)	(134)	(Std.)	(77)

Table 2: Distribution of essay scores, and average (std) of essay length (in words), Train data.

2.1 Set 1

Dataset 1 is comprised of essays written in 2012 and 2013 as part of a large-scale college-level examination in the United States, by a mix of native and non-native speakers of English. The essays respond to a "criticize an argument" task, where a test-taker is given a short prompt text of about 150 words that typically describes a setting where some recommendation is made or a claim is put forward. The task of the test-taker is then to critically evaluate the arguments presented in support of the claim. An example prompt is shown in Figure 1.

2.2 Set 2

The second dataset is used for evaluating the generalization of the text-topicality models to a different type of essays. Essays in this dataset are written in

says that in both sentences, half the content words are topical, whereas the type-based model says that only 1 out of 5 different content words is topical in the first sentence, and 4 out of 8 in the second.

> In surveys Mason City residents rank water sports (swimming, boating and fishing) among their favorite recreational activities. The Mason River flowing through the city is rarely used for these pursuits, however, and the city park department devotes little of its budget to maintaining riverside recreational facilities. For years there have been complaints from residents about the quality of the river's water and the river's smell. In response, the state has recently announced plans to clean up Mason River. Use of the river for water sports is therefore sure to increase. The city government should for that reason devote more money in this year's budget to riverside recreational facilities.
>
> Write a response in which you examine the stated and/or unstated assumptions of the argument. Be sure to explain how the argument depends on the assumptions and what the implications are if the assumptions prove unwarranted.

Figure 1: An example Set 1 prompt.

a more open-ended "support your position on an argument" genre, where the prompt is typically a single sentence that puts forward a general claim, such as "As people rely more and more on technology to solve problems, the ability of humans to think for themselves will surely deteriorate." This task is administered on the same test as the one discussed above, and the general properties, such as scale and distribution of scores, are similar.

2.3 TOEFL Set

The third dataset will be used to assess generalization of the findings regarding text-topicality models to shorter essays written by non-native speakers of English – a generally less English-proficient population than writers in Sets 1 and 2. This dataset is publicly available from the Linguistic Data Consortium (Blanchard et al., 2013).[2] This set contains 12,100 essays written for the Test of English as a Foreign Language (TOEFL), responding to the question "Do you agree or disagree with the following statement?", a genre similar to that of Set 2. The essays in this set were written in response to 8 different prompts, such as: "A teacher's ability to relate well with students is more important than excellent knowledge of the subject being taught." Only coarse-grained scores are provided, corresponding to low, medium, and high proficiency, which we represent as scores 1, 2, and 3, respectively. The data was partitioned so that 500 essays per prompt are used in Dev to estimate the topical lists; the remaining essays are split 75% (Train) and 25% (Test) within each prompt.

3 Models of Word Topicality

Let T_1 ... T_m be sets of essays responding to m different prompts t_1 ... t_m. For a word w and a prompt t_k, we define the following contingency table of counts, where $\neg w$ corresponds to any content word other than w and $\neg T_k$ corresponds to $\cup_{r \neq k} T_r$:

	T_k	$\neg T_k$
w	A_{11}	A_{12}
$\neg w$	A_{21}	A_{22}

We define the following word topicality models. The first model, **LH**, due to Lin and Hovy (2000), quantifies the topicality of a word in a topic as a reduction in the entropy of topic distribution achieved by partitioning on the word w, scaled so that the resulting value is distributed according to χ^2. Note that to avoid division by zero when $1 - p_1 = 0$, we only consider words with $A_{12} > 0$.

$$LH_{w,k} = -2log \frac{p^{A_{11}+A_{21}}(1-p)^{A_{12}+A_{22}}}{p_1^{A_{11}}(1-p_1)^{A_{12}} p_2^{A_{21}}(1-p_2)^{A_{22}}} \tag{1}$$

where the proportions p, p_1, and p_2 are given by:

$$p = \frac{A_{11} + A_{21}}{A_{11} + A_{21} + A_{12} + A_{22}} \tag{2}$$

$$p_1 = \frac{A_{11}}{A_{11} + A_{12}}; \quad p_2 = \frac{A_{21}}{A_{21} + A_{22}} \tag{3}$$

From this definition, we derive three word topicality weights – the first using the continuous values of topicality mapped to the $[0, 1]$ range, the second – binarized to separate out only words that reach the 0.001 significance, the third – a binarized model

[2] LDC Catalogue No: LDC2014T06

with a more permissive threshold for 0.05 significance, which would create larger but noisier sets of topical vocabulary.[3]

$$\alpha_1^{LH}(w,k) = \frac{LH(w,k)}{max_{v \in T_k} LH(v,k)} \tag{4}$$

$$\alpha_2^{LH}(w,k) = \begin{cases} 1 & \text{if } LH(w,k) > 10.83 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

$$\alpha_3^{LH}(w,k) = \begin{cases} 1 & \text{if } LH(w,k) > 3.84 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

The second model, **HBA**, due to Higgins et al. (2006), quantifies topicality of a word as the geometric mean of its probability of occurrence in the topic and the complement of its probability of occurrence overall. Thus, the more topical words tend to occur more frequently in the current topic and more rarely in general (this reasoning is similar to tfidf). According to this model, the weight of a word in a topic is defined as follows:[4]

$$\alpha^{HBA} = \sqrt{\frac{A_{11}}{A_{11}+A_{21}} \times \frac{A_{21}+A_{22}}{A_{11}+A_{21}+A_{12}+A_{22}}} \tag{7}$$

Lastly, we define a simple (**S**) cutoff-based binary index, where the word is topical if it is likelier to occur in the current topic than overall:

$$\alpha^S = \begin{cases} 1 & \text{if } \frac{A_{11}}{A_{11}+A_{21}} > \frac{A_{11}+A_{12}}{A_{11}+A_{21}+A_{12}+A_{22}} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

4 Models of Text Topicality

For an essay e, let Y be a set of all content word[5] types in e and let O be a set of all content word tokens.[6] Further, let α_w be the topicality value of the word $w \in e$. We then define text topicality as the proportion of topical words (for binary word topicality indices) or mean topicality per word (for continuous word topicality indices), for types and tokens, as follows:

$$TypeTop(e) = \frac{1}{|Y|} \sum_{w \in Y} \alpha_w \tag{9}$$

$$TokTop(e) = \frac{1}{|O|} \sum_{w \in O} \alpha_w \tag{10}$$

We observe that all the word topicality models defined above essentially produce a final list of topical words, based on some estimation set. This introduces a dependence between text length and its topicality, especially under the type-based definition: The longer the text, the less likely it is that the next new word would be topical – simply because there are only so many topical words, and their supply diminishes with every newly chosen word, whereas the (theoretical) supply of non-topical words is infinite. This reasoning suggests that the longer the text, the less topical it would be, on average. Note that we are not implying that longer texts digress more; it is just that the modeling of topicality that is based on a finite list of topical words is inherently biased against longer essays.

Figure 2 shows that this is indeed the case, using a separate set of 3,000 essays responding to the same task as Set 1, using the α^S word topicality index and type-based aggregation. The different series correspond to sets of essays within a certain length band, with color codes ranging from the lightest blue for the shortest essays (shorter than 2 standard deviations below mean length) to the lightest orange for the longest ones (more than 1.5 standard deviation longer than mean length). It is clearly the case that longer essays tend to be less topical, as moving from blue to red to orange generally aligns with moving down the topicality axis. Thus, given that essay length is typically strongly positively correlated with essay scores, we expect that topicality would be negatively correlated with score. However, separating essays by length bands reveals that the relationship between topicality and score is in fact *positive* – when length is held approximately constant, better essays tend to be more topical.[7] These obser-

[3] For all indices, we set the value to 0 if $p_2 \geq p_1$, even though a reduction in entropy due to a partition on w could occur when the topic is substantially *less* likely given that w occurred.

[4] In Higgins et al. (2006), the probability of occurrence in general is estimated from a different dataset than that used to estimate prompt-specific probabilities. However, presumably, the general dataset would contain some number of essays responding the current prompt, so we believe our approximation is faithful to the spirit of the original.

[5] We assume function words are irrelevant for topicality.

[6] tYpes vs tOkens

[7] Observe the upward slope of each series in Figure 2.

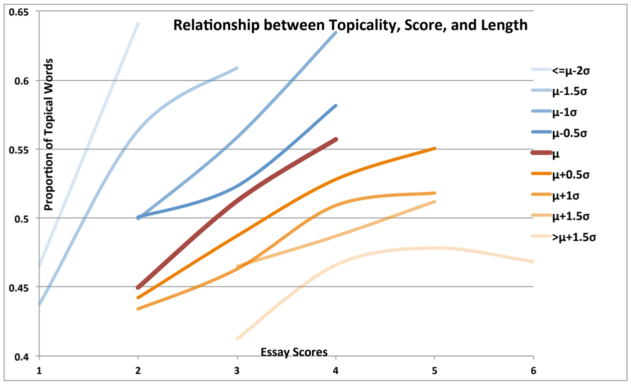

Figure 2: Illustration of the relationship between essay score, essay length, and topicality, using α^S index in type aggregation, using an additional sample of 3,000 essays responding to the task in Set 1. The series correspond to length bands, with the lightest blue line showing mean topicality per score level for essays that are more than 2 standard deviations below mean essay length.

vations suggest that the estimated topicality of the essay needs to be scaled to compensate for "baseline" topicality differences that are due to length. We therefore define a length-scaled version of the two indices as follows:

$$TypeTopL(e) = \frac{log(|Y|)}{|Y|} \sum_{w \in Y} \alpha_w \qquad (11)$$

$$TokTopL(e) = \frac{log(|O|)}{|O|} \sum_{w \in O} \alpha_w \qquad (12)$$

5 Selecting Best Topicality Model(s)

We evaluate each of the 5 word-topicality models (α) with each of the 4 text-aggregation methods (types/tokens, scaled/unscaled) – 20 models in total – for their ability to predict essay score above and beyond the prediction based on essay length. Essay length is a well-known confounder for essay scoring systems (Page, 1966): It is a strong predictor of essay score (r=0.65 for Set 1); yet, an automated essay scoring system needs to capture additional aspects of essay quality construct beyond the basic English production fluency captured by essay length. Our measure of success is therefore **partial correlation** r_p between the feature and the human-provided essay score, excluding the effect of essay length.[8] Table 3 shows the results.

We make the following observations based on these results.

First, the relative merits of various topicality models generalize very well across the three sets. We calculated rank-order (Spearman) correlations between the 20 partial correlations for the various models on the three pairs of datasets. Thus, the rank order correlation between column r$_p$ for Set 1 and column r$_p$ for Set 2 in Table 3 is ρ = 0.92; Set 1 vs TOEFL ρ = 0.93; Set 2 vs TOEFL ρ = 0.98.

Second, it is clearly the case that the text topicality indices based on continuous word topicality indices (LH$_1$, HBA) are less effective, their partial correlations with score excluding length being within 0.15 band around zero (lines 1-8 in Table 3). Although some overall correlations with score are reasonable (such as 0.262 in line 4, Set 1; 0.235 in line 2, Set 1), these are mostly accounted for by the even higher correlation with essay length. This suggests that accounting for the nuances of the extent of the topicality of each word is generally not effective – once the word is topical enough, it matters not just how topical it is. Or, at the very least, we have not yet found a way to devise an effective continuous topicality score for a word.

Let us now consider the more effective cutoff-based binary indices (LH$_2$, LH$_3$, SIMPLE), and evaluate the effects of the two manipulations applied across the word topicality models: the log scaling and the use of types vs tokens.

Log Scaling: This manipulation is effective in every single case (compare odd lines n to even lines n+1 for $n > 8$, for each of the datasets, for a total of 18 comparisons).

Type vs Token: Types are better than tokens in every single case (compare lines n to lines n+2 within each word topicality model, for $n > 8$, for each of the datasets, for a total of 18 comparisons).

[8] Since some of the indices are scaled by log length, we calculated second-order partial correlations excluding the linear effects of both length and log-length on Set 1. The resulting values were very close to the first-order partial correlation values that control for length only, and did not change the comparative standings of the various models. For simplicity, Table 3 reports first-order partial correlations controlling for length for all models.

ID	Word	Tok/	Log	Set 1			Set 2			TOEFL		
	Model	Type	Scale?	R$_s$	R$_l$	r$_p$	R$_s$	R$_l$	r$_p$	R$_s$	R$_l$	r$_p$
1	HBA	tok	no	-.014	-.114	.079	-.194	-.217	-.070	-.208	-.216	-.102
2	HBA	tok	yes	.235	.248	.099	.020	.077	-.041	-.041	.040	-.080
3	HBA	typ	no	.114	.152	.020	-.094	-.028	-.101	-.163	-.076	-.146
4	HBA	typ	yes	.262	.358	.042	.055	.169	-.077	-.050	.087	-.125
5	LH$_1$	tok	no	-.033	-.102	.044	-.138	-.194	-.014	-.124	-.185	-.020
6	LH$_1$	tok	yes	.092	.076	.056	-.037	-.059	.002	-.037	-.052	-.008
7	LH$_1$	typ	no	.051	.057	.018	-.096	-.095	-.045	-.117	-.106	-.068
8	LH$_1$	typ	yes	.148	.190	.033	-.010	.015	-.026	-.044	-.005	-.050
9	LH$_2$	tok	no	.049	-.103	.154	.019	-.014	.150	-.054	-.141	.035
10	LH$_2$	tok	yes	.304	.269	.176	.235	.155	.178	.139	.155	.061
11	LH$_2$	typ	no	-.096	-.317	.152	-.022	-.255	.202	-.072	-.224	.074
12	LH$_2$	typ	yes	.135	-.024	.199	.197	.006	.257	.146	.060	.137
13	LH$_3$	tok	no	.070	-.092	.171	.042	-.126	.169	-.016	-.112	.062
14	LH$_3$	tok	yes	.343	.310	.195	.283	.206	.200	.199	.219	.090
15	LH$_3$	typ	no	-.052	-.279	.177	.013	-.223	.220	-.016	-.177	.109
16	LH$_3$	typ	yes	.207	.054	.227	.261	.079	.280	.228	.143	.179
17	SIMPLE	tok	no	.105	-.073	.201	.086	-.098	.202	.078	-.045	.129
18	SIMPLE	tok	yes	.430	.418	.228	.385	.324	.240	.338	.368	.163
19	SIMPLE	typ	no	.016	-.219	.214	.080	-.165	.256	.106	-.070	.181
20	SIMPLE	typ	yes	.349	.227	.272	.396	.237	.328	.399	.334	.267

Table 3: Performance of the different word-topicality models (α), with or without log length scaling, in type or token aggregation, on the three datasets, in terms of Pearson correlation with essay score (R$_s$), Pearson correlation with essay length (R$_l$), and partial correlation with score controlling for length (r$_p$). Evaluations are performed on Train data in each dataset.

Finally, we observe that among the cutoff models, the more permissive, the better – the model with a stricter significance threshold for topicality performs worse than the one with a looser threshold, which in turn performs worse than a simple cutoff model with no significance test at all (compare line n to line n+4, for $n > 8$, in Table 3). This suggests that richer but noisier topical lists are generally more effective, in the essay scoring context.

Following these observations, we select the type-aggregated log-scaled simple topicality index for evaluation within an essay scoring system for the three datasets.

6 Essay Scoring Experiments

In this section, we present an evaluation of the best topicality index for each of the three datasets as a feature in a comprehensive, state-of-art essay scoring system. The baseline engine (e-rater®, described in Burstein et al. (2013)) computes more than 100 micro-features, which are aggregated into macro-features aligned with specific aspects of the writing construct. The system incorporates macro-features measuring grammar, usage, mechanics, organization, development, etc; Table 4 shows the nine macro-features, with examples of micro-features. In addition, we put essay length (number of words) as the 10th macro-feature into the baseline model, to ascertain that any gains observed in the experimental condition are not due to the introduction of length as part of the scaling in the topicality feature.

In the **baseline** condition, a scoring model is built over the ten macro-features using linear regression on the Train set and evaluated on the Test set, for each of the datasets. In the **experimental** condition, the topicality index is added as the 11th macro-feature into the linear regression model; the experimental system is also trained on Train set and evaluated on Test set, for each of the datasets. We evaluate essay scoring performance using Pearson correlation with human holistic score.

To test statistical significance of the improvements, we use Wilcoxon signed-rank test for matched pairs. We calculate the baseline and experimental performance on each prompt separately, and use the 76 pairs of values (for each of Sets 1 and 2) and 8 pairs of values (for TOEFL) as inputs for the test. We use VassarStats for performing the significance tests.[9] Table 5 show the results.

Macro- Features	**Example Micro-Features**
Grammar	garbled , run-on, fragmented sentences
Usage	determiner-noun agreement errors, noun number errors, missing article
Mechanics	spelling, capitalization, punctuation errors
Organization	use of discourse elements, such as thesis, support, conclusion
Development	size of discourse elements
Vocabulary 1	av. word frequency
Vocabulary 2	av. word length
Idiomaticity	use of appropriate prepositions, use of collocational patterns
Sentence Variety	use of sentences with various levels of syntactic complexity
Length	number of words in the essay

Table 4: Baseline essay scoring system (9 macro-features from a state-of-art essay scoring system, and essay length).

We find that the addition of the topicality feature leads to a statistically significant improvement over the baseline for each of the three datasets. In an additional set of experiments, we removed essay length from both the baseline and the experimental conditions to check whether the topicality feature would improve upon a state-of-art essay scoring system as-is; we found an improvement in all the three datasets, at the same significance levels as those reported in Table 5.

7 Related Work

The two approaches that are most closely related to the current work are those of Higgins et al. (2006) and Lin and Hovy (2000), who present word topicality models based on a comparison between the distribution of words in on-topic and off-topic texts. Indeed, these models were the starting point of our work, along with a simpler comparison model based on raw frequencies. Higgins et al. (2006) aggregated the word-level scores using unscaled token-level ag-

[9] http://vassarstats.net/

Data	Performance (r)		Signif.
	Baseline	Experimental	Level
Set 1	0.762	0.766	0.0001
Set 2	0.800	0.804	0.0001
TOEFL	0.747	0.749	0.05

Table 5: Results of essay scoring experiments. Performance is reported in terms of Pearson correlation with human essay score. 2-tailed p-values for rejecting the null hypothesis of no improvement are shown in the last column. The Wilcoxon test statistics are: W= 1486, n = 73 (Set 1); W = 2064, n = 71 (Set 2); W = 25, n = 7 (TOEFL).

gregation; our results suggest that this aggregation method can be improved upon by log-length scaling and type-based aggregation. We also showed that Lin and Hovy (2000) topicality models produce better predictions of essay quality, with appropriate scaling and aggregation.

Louis and Higgins (2010) and Higgins et al. (2006) address the task of detecting off-topic essays without on-topic training materials. Persing and Ng (2014) reported a study where essays were scored on an analytic rubric of adherence to the prompt; while this is a promising way to evaluate text-topicality models intrinsically, the reliability of the annotations was low (r=0.234). Content scoring was also studied for essays written in response to an extensive reading or listening prompt – quality of content is then related to integrating information from the source materials (Beigman Klebanov et al., 2014; Kakkonen et al., 2005; Lemaire and Dessus, 2001).

A related direction of research implicitly treats topicality as a part of a more generalized notion of "good content," namely, words that are used by good writers. The approach to estimating the quality of content is to compare the content of the current text to sets of training texts that represent various score points (Attali and Burstein, 2006; Kakkonen et al., 2005; Xie et al., 2012). In this approach, there is no differentiation between content that is topical and other words that might be used for other reasons, such as discourse markers used for organizational purposes or spurious, shell-like elements (Madnani et al., 2012); an essay that is dissimilar from high-scoring essays on all or some of these accounts is likely to be viewed as having "bad content." An essay rife with misspellings would likewise be seen as having "bad content", because the model high-scoring essays are generally not prone to misspellings. In contrast, our topicality lists are estimated based on a random sample of essays, including low scoring essays; this allows introduction of common misspellings of words frequently used to address the given topic into the topical lists. For example, one of the topical lists includes more than a dozen misspellings of the word *contemporaries*.

There is a large body of work using topic models to capture different topics typically addressed in a corpus of text (Mimno et al., 2011; Newman et al., 2011; Gruber et al., 2007; Blei et al., 2003). In this general framework, each text can address a few different topics and the number and identity of topics for the corpus is typically unknown. In our setting, we assume that each essay is on a single topic, and that topic is known in advance.[10] However, many of these topics are very open-ended, so they might exhibit non-trivial sub-topical structures. For example, a topic about cultural role models might be dealt with by discussing politicians, musicians, sportsmen – each of these could yield a specific sub-topic. In fact, Persing and Ng (2014) used LDA to create sub-topics in this way, and derived features to predict prompt-adherence of an essay. The authors found that in order to make these features more effective, it was beneficial for humans to go over the topics and assign relevance estimates for each sub-topic.

8 Conclusion

In this paper, we addressed the problem of quantifying the overall extent to which a test-taker's essay deals with the topic it is assigned (prompt). We experimented with a number of approaches for quantifying the topicality of a word, and with a number of approaches for aggregating word-level topicality into text-level topicality. We found that type-based, log-length scaled aggregation generally works better than the token-based and unscaled one, for the task of predicting the holistic quality of essays. The findings of the effectiveness of log length scaling and of type-based accounting when estimating the topicality of an essay for the purposes of holistic scoring are novel contributions of this work.

[10] The high-stakes nature of the examination ensures that these assumptions are rarely wrong.

We also showed that incorporation of text-topicality into essay scoring yields a significant improvement for two different writing tasks over a very strong baseline – a state-of-art essay scoring system augmented with an essay length feature. A significant improvement is also observed on the publicly available set of TOEFL essays, even though the set is smaller, there are only a handful of different prompts, the essays are shorter and less proficiently written, and the scores are given on a coarser-grained scale than for the other two datasets. The demonstration of the excellent generalization of the relative merits of the various topicality models across three datasets and the effectiveness of the topicality feature for improving essay scoring on the three sets is another novel contribution of this work; it suggests robustness of our findings regarding the relationship between topicality, length and quality of essays.

An interesting direction of future work is an intrinsic evaluation of topicality indices against human judgments of topicality. This is a difficult annotation task (Persing and Ng, 2014), and, to our knowledge, no reliable protocol exists for this task.

References

Yigal Attali and Jill Burstein. 2006. Automated Essay Scoring With e-rater®V.2. *Journal of Technology, Learning, and Assessment*, 4(3).

Beata Beigman Klebanov, Nitin Madnani, Jill Burstein, and Swapna Somasundaran. 2014. Content importance models for scoring writing from sources. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 247–252, Baltimore, Maryland, June.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. TOEFL11: A corpus of non-native english. Educational Testing Service Research Report No. RR-13-24.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Jill Burstein, Joel Tetreault, and Nitin Madnani. 2013. The E-rater®Automated Essay Scoring System. In M. Shermis and J. Burstein, editors, *Handbook of Automated Essay Scoring: Current Applications and Future Directions*. New York: Routledge.

Peter Foltz, Walter Kintsch, and Thomas Landauer. 1998. The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, 25(2):285–307.

Amit Gruber, Yair Weiss, and Michal Rosen-Zvi. 2007. Hidden topic markov models. *Journal of Machine Learning Research - Proceedings Track*, 2:163–170.

Derrick Higgins, Jill Burstein, Daniel Marcu, and Claudia Gentile. 2004. Evaluating multiple aspects of coherence in student essays. In *Proceedings of NAACL*, pages 185–192, Boston, Massachusetts, USA, May.

Derrick Higgins, Jill Burstein, and Yigal Attali. 2006. Identifying off-topic student essays without topic-specific training data. *Natural Language Engineering*, 12(2):145–159.

Tuomo Kakkonen, Niko Myller, Jari Timonen, and Erkki Sutinen. 2005. Automatic Essay Grading with Probabilistic Latent Semantic Analysis. In *Proceedings of the Second Workshop on Building Educational Applications Using NLP*, pages 29–36, Ann Arbor, Michigan, June.

Benoît Lemaire and Philippe Dessus. 2001. A System to Assess the Semantic Content of Student Essays. *Journal of Educational Computing Research*, 24:305–320.

Chin-yew Lin and Eduard Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *Proceedings of COLING*, pages 495–501.

Annie Louis and Derrick Higgins. 2010. Off-topic essay detection using short prompt texts. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 92–95, Los Angeles, California, June.

Nitin Madnani, Michael Heilman, Joel Tetreault, and Martin Chodorow. 2012. Identifying high-level organizational elements in argumentative discourse. In *Proceedings of NAACL*, pages 20–28.

David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 262–272, Stroudsburg, PA, USA.

David Newman, Edwin Bonilla, and Wray Buntine. 2011. Improving topic coherence with regularized topic models. In J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 496–504.

Ellis B. Page. 1966. The Imminence of Grading Essays by Computer. *Phi Delta Kappan*, pages 238–243.

Isaac Persing and Vincent Ng. 2014. Modeling prompt adherence in student essays. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1534–1543, Baltimore, Maryland, June.

Swapna Somasundaran, Jill Burstein, and Martin Chodorow. 2014. Lexical chaining for measuring discourse coherence quality in test-taker essays. In *Proceedings of COLING*, pages 950–961.

Joel Tetreault, Daniel Blanchard, Aoife Cahill, and Martin Chodorow. 2012. Native tongues, lost and found: Resources and empirical evaluations in native language identification. In *Proceedings of COLING 2012*, pages 2585–2602, Mumbai, India, December.

Shasha Xie, Keelan Evanini, and Klaus Zechner. 2012. Exploring content features for automated speech scoring. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 103–111, Montréal, Canada, June.

Su-Youn Yoon and Shasha Xie. 2014. Similarity-based non-scorable response detection for automated speech scoring. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 116–123, Baltimore, Maryland, June.