Content Importance Models for Scoring Writing From Sources

Beata Beigman Klebanov **Nitin Madnani** **Jill Burstein** **Swapna Somasundaran**
Educational Testing Service
660 Rosedale Road
Princeton, NJ 08541
{bbeigmanklebanov,nmadnani,jburstein,ssomasundaran}@ets.org

Abstract

Selection of information from external sources is an important skill assessed in educational measurement. We address an integrative summarization task used in an assessment of English proficiency for non-native speakers applying to higher education institutions in the USA. We evaluate a variety of content importance models that help predict which parts of the source material should be selected by the test-taker in order to succeed on this task.

1 Introduction

Selection and integration of information from external sources is an important academic and life skill, mentioned as a critical competency in the Common Core State Standards for English Language Arts/Literacy: College-ready students will be able to "gather relevant information from multiple print and digital sources, assess the credibility and accuracy of each source, and integrate the information while avoiding plagiarism."[1]

Accordingly, large-scale assessments of writing incorporate tasks that test this skill. One such test requires test-takers to read a passage, then to listen to a lecture discussing the same topic from a different point of view, and to summarize the points made in the lecture, explaining how they cast doubt on points made in the reading. The quality of the information selected from the lecture is emphasized in excerpts from the scoring rubric for this test (below); essays are scored on a 1-5 scale:

Score 5 successfully selects the important information from the lecture and coherently and accurately presents this information in relation to the relevant information presented in the reading.

Score 4 is generally good in selecting the important information from the lecture ..., but it may have a minor omission.

Score 3 contains some important information from the lecture ..., but it may omit one major key point.

Score 2 contains some relevant information from the lecture ... The response significantly omits or misrepresents important points.

Score 1 provides little or no meaningful or relevant coherent content from the lecture.

The ultimate goal of our project is to improve automated scoring of such essays by taking into account the extent to which a response integrates important information from the lecture. This paper reports on the first step aimed at automatically assigning importance scores to parts of the lecture. The next step – developing an essay scoring system using content importance models along with other features of writing quality, will be addressed in future work. A simple essay scoring mechanism will be used for evaluation purposes in this paper, as described in the next section.

2 Design of Experiment

In evaluations of summarization algorithms, it is common practice to derive the gold standard content importance scores from human summaries, as done, for example, in the pyramid method, where the importance of a content element corresponds to the number of reference human summaries that make use of it (Nenkova and Passonneau, 2004). Selection of the appropriate content plays a crucial role in attaining a high score for the essays we consider here, as suggested by the quotes from the scoring rubric in §1, as well as by a corpus study by Plakans and Gebril (2013). We therefore observe that high-scoring essays can be thought

[1] http://www.corestandards.org/ELA-Literacy/CCRA/W.

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 247–252,
Baltimore, Maryland, USA, June 23-25 2014. ©2014 Association for Computational Linguistics

of as high-quality human summaries of the lecture, albeit containing, in addition, references to the reading material and language that contrasts the different viewpoints, making them a somewhat noisy gold standard. On the other hand, since low-scoring essays contain deficient summaries of the lecture, our setup allows for a richer evaluation than typical in studies using gold standard human data only, in that a good model should not only agree with the gold standard human summaries but should also disagree with sub-standard human summaries. We therefore use correlation with essay score to evaluate content importance models.

The evaluation will proceed as follows. Every essay E is responding to a test prompt that contains a lecture L and a reading R. We identify the essay's overlap with the lecture:

$$O(E, L) = \{x | x \in L, x \in E\} \qquad (1)$$

where the exact definition of x, that is, what is taken to be a single unit of information, will be one of the parameters to be studied. The essay is then assigned the following score by the content importance model M:

$$S_M(E) = \frac{\Sigma_{x \in O(E,L)} w_M(x) \times C(x, E)}{n_E} \qquad (2)$$

where $w_M(x)$ is the importance weight assigned by model M to item x in the lecture, $C(x, E)$ is the count of tokens in E that realize the information unit x, and n_E is the number of tokens in the essay. In this paper, the distinction between x and C is that between type and token count of instances of that type.[2] This simple scoring mechanism quantifies the rate of usage of important information per token in the essay. Finally, we calculate the correlation of scores assigned to essays by model M with scores assigned to the same essays by human graders.

This design ensures that once x is fixed, all the content importance models are evaluated within the same scoring scheme, so any differences in the correlations can be attributed to the differences in the weights assigned by the importance models.

3 Content Importance Models

Our setting can be thought of as a special kind of summarization task. Test-takers are required to summarize the lecture while referencing the reading, making this a hybrid of single- and multi-document summarization, where one source is treated as primary and the other as secondary.

We therefore consider models of content importance that had been found useful in the summarization literature, as well as additional models that utilize a special feature of our scenario: We have hundreds of essays of varying quality responding to any given prompt, as opposed to a typical news summarization scenario where a small number of high quality human summaries are available for a given article. A sample of these essays can be used when developing a content importance model.

We define the following importance models. For all definitions, x is a unit of information in the lecture; $C(x, t)$ is the number of tokens in text t that realize x; n_L and n_R are the number of tokens in the lecture and the reading, respectively.[3]

Naïve: $w(x) = 1$. This is a simple overlap model.

Prob: $w(x) = \frac{C(x,L)}{n_L}$, an MLE estimate of the probability that x appears in the lecture. Those x that appear more are more important.

Position: $w(x) = \frac{FP(x)}{n_L}$, where $FP(x)$ is the offset of the first occurrence of x in the lecture. The offset corresponds to the token's serial number in the text, 1 through n_L.

LectVsRead: $w(x) = \frac{C(x,L)}{n_L} - \frac{C(x,R)}{n_R}$, that is, the difference in the probabilities of occurrence of x in the lecture and in the reading passage that accompanies the lecture. This model attempts to capture the contrastive aspect of importance – the content that is unique to the lecture is more important than the content that is shared by the lecture and the reading.

The following two models capitalize on evidence of use of information in better and worse essays. For estimating these models, we sample, for each prompt, a development set of 750 essays responding to the prompt (that is, addressing a given pair of lecture and reading stimuli). Out of these, we take, for each prompt, all essays at score points

[2] In the future, we intend to explore more complex realization functions, allowing paraphrase, skip n-grams (as in ROUGE (Lin, 2004)), and other approximate matches, such as misspellings and inflectional variants.

[3] Prob, Position, and LectVsRead models normalize by n_R and n_L to enable comparison of essays responding to different lecture + reading stimuli (prompts).

4 and 5 (**EGood**) and all essays at score points 1 and 2 (**EBad**). These data do not overlap with the experimental data described in section 4. In both definitions below, e is an essay.

Good: $w(x) = \frac{|\{e \in EGood | x \in e\}|}{|EGood|}$. An x is more important if more good essays use it. Hong and Nenkova (2014) showed that a variant of this measure used on pairs of articles and their abstracts from the New York Times effectively identified words that typically go into summaries, *across topics*. In contrast, our measurements are prompt-specific.

GoodVsBad: $w(x) = \frac{|\{e \in EGood | x \in e\}|}{|EGood|} - \frac{|\{e \in EBad | x \in e\}|}{|EBad|}$. An x is more important if good essays use it more than bad essays. To our knowledge, this measure has not been used in the summarization literature, probably because a large sample of human summaries of varying quality is typically not available.

4 Data

We use 116 prompts drawn from an assessment of English proficiency for non-native speakers. Each prompt contains a lecture and a reading passage. For each prompt, we sample about 750 essays. Each essay has an operational score provided by a human grader. Table 1 shows the distribution of essay scores; mean score is 3. Text transcripts of the lectures were used.

Score	1	2	3	4	5
Proportion	0.13	0.18	0.35	0.25	0.09

Table 1: Distribution of essay scores.

5 Results

Independently from the content importance models, we address the effect of the granularity of the unit of information. Intuitively, since all the materials for a given prompt deal with the same topic, we expect large unigram overlaps between lecture and reading, and between good and bad essays, whereas n-grams with larger n can be more distinctive. On the other hand, larger n lead to misses, where an information unit would fail to be identified in an essay due to a paraphrase, thus impairing the ability of the scoring function to use the content importance model effectively. We therefore evaluate each content importance model for different granularities of the content unit x: n-grams for $n = 1, 2, 3, 4$. Table 2 shows the correlations with essay scores.

Content Importance Model	Pearson's r			
	n=1	n=2	n=3	n=4
Naïve	*0.24*	0.27*	0.24	0.20
Prob	0.04	0.14	0.17	0.14
Position	0.22	**0.30***	0.26*	0.20
LectVsRead	0.09	0.25*	**0.31***	0.26*
Good	0.07	0.15	0.10	0.07
GoodVsBad	**0.54***	**0.42***	**0.32***	0.21

Table 2: Correlations with essay scores attained by content models, for various definitions of information unit (n-grams with $n = 1, 2, 3, 4$). Five top scores are boldfaced. The baseline performance is shown in underlined italics. Correlations that are significantly better ($p < 0.05$) than the naïve $n = 1$ model are marked with an asterisk. We use McNemar (1955, p. 148) test for significance of difference between same-sample correlations. $N = 85,252$ for all correlations.

6 Discussion

The Naïve model with $n = 1$ can be considered a baseline, corresponding to unweighted word overlap between the lecture and the essay. This model attains a significant positive correlation with essay score ($r = 0.24$), suggesting that, in general, better writers use more material from the lecture.

Our next observation is that the Prob and Good models do not improve over the baseline, that is, their weighting schemes generally assign higher weights to the wrong units. We believe the reason for this is that the most highly used n-grams, in the lecture and in the essays, correspond to general topical and functional elements. The importance of these elements is discounted in the more effective Position, LectVsRead, and GoodVsBad models, highlighting subtler aspects of the lecture.

Next, let us consider the granularity of the units of information. We observe that 4-grams are inferior to trigrams for all models, suggesting that data sparsity is becoming a problem for matching 4-word sequences. For models that assign weight based on one or two sources (lecture, or lecture and reading) – Naïve, Position, LectVsRead – unigram models are generally ineffective, while bigram and trigram models significantly outperform the baseline. We interpret this as suggesting that it is certain particular, detailed aspects of the topical concepts that constitute the important nuggets in the lecture; these are usually realized by multiword sequences.

The GoodVsBad models show a different pattern, obtaining the best performance with a unigram version. These models are sensitive to data sparsity not only when matching essays to the lecture (this problem is common to all models) but also during model building. Recall that the weights in a GoodVsBad model are estimated based on differential use in samples of good and bad essays. The estimation of use-in-a-corpus is more accurate for smaller n, because longer n-grams are more susceptible to paraphrasing, which leads to under-estimation of use. Assuming that paraphrasing behavior of good and bad writers is not the same – in fact, there is corpus evidence that better writers paraphrase more (Burstein et al., 2012) – the resulting inaccuracies might impact the estimation of differential use in a systematic manner, making the $n > 1$ models less effective than the unigrams. Given that (a) the GoodVsBad bigram model is the second best overall in spite of the shortcomings of the estimation process, and (b) that the bigram models worked better than unigram models for all the other content importance models, the GoodVsBad bigram model could probably be improved significantly by using a more flexible information realization mechanism.

To illustrate the information assigned high importance by different models, consider a lecture discussing advantages of fish farming. The top-scoring Good bigrams are topical expressions (*fish farming*), functional bigrams around *fish* and *farming*,[4] aspects of content dealt with at length in the lecture (*wild fish*, *commercial fishing*), bigrams referencing some of the claims – fish containing *less fat* and being used for *fish meal*. In addition, this model picks out some sequences of function words and punctuation (*of the*, *are not*, *", and"*, *", the"*) that suggest that better essays tend to give more detail (hence have more complex noun phrases and coordinated constructions) and to draw contrast.

For the bigram GoodVsBad model, the topical bigram *fish farming* is not in the top 20 bigrams. Although some bigrams are shared with the Good model, the GoodVsBad model selects additional details about the claims, such as the contrast between *inedible fish* and *edible fish* that is *eaten by humans*, as well as reference to *chemicals used* in farming and to the claim that wild fish are *already endangered* by other practices.

The most important bigrams according to the LectVsRead model include functional bigrams around *fish* and *farming*, functional sequences (*that the*, *is a*), as well as *commercial fishing* and *edible fish*. Also selected are functional bigrams around *consumption* and *species*, hinting, indirectly, at the edibility differences between species. Finally, this model selects almost all bigrams in *the reading passage makes*, *the reading makes claims that* and *the reading says*. While distinguishing the lecture from the reading, these do not capture topic-relevant content of the lecture.

The GoodVsBad unigram model selects *poultry*, *endangered*, *edible*, *chemicals* among its top 6 unigrams,[5] effectively touching upon the connection with other farm-raised foods (*poultry*, *chemicals*), with wild fish (*endangered*) and with human benefit (*edible*) that are made in the lecture.

7 Related work

Modern essay scoring systems are complex and cover various aspects of the writing construct, such as grammar, organization, vocabulary (Shermis and Burstein, 2013). The quality of content is often addressed by features that quantify the similarity between the vocabulary used in an essay and reference essays from given score points (Attali and Burstein, 2006; Foltz et al., 2013; Attali, 2011). For example, Attali (2011) proposed a measure of differential use of words in higher and lower scoring essays defined similarly to GoodVsBad, without, however, considering the source text at all. Such features can be thought of as content quality features, as they implicitly assume that writers of better essays use better content. However, there are various kinds of better content, only one of them being selection of important information from the source; other elements of content originate with the writer, such as examples, discourse markers, evaluations, introduction and conclusion, etc. Our approach allows focusing on a particular aspect of content quality, namely, selection of appropriate materials from the source.

[4] such as *that fish*, *of fish*, *farming is*, *", fish"*

[5] the other two being *fishing* and *used*.

Our results are related to the findings of Gurevich and Deane (2007) who studied the difference between the reading and the lecture in their impact on essay scores for this test. Using data from a single prompt, they showed that the difference between the essay's average cosine similarity to the reading and its average cosine similarity to the lecture is predictive of the score for non-native speakers of English, thus using a model similar to LectVsRead, although they took all lecture, reading, and essay words into account, in contrast to our model that looks only at n-grams that appear in the lecture. Our study shows that the effectiveness of lecture-reading contrast models for essay scoring generalizes to a large set of prompts. Similarly, Evanini et al. (2013) found that overlap with material that is unique to the lecture (not shared with the reading) was predictive of scores in a spoken source-based question answering task.

In the vast literature on summarization, our work is closest to Hong and Nenkova (2014) who studied models of word importance for multi-document summarization of news. The Prob, Position, and Good models are inspired by their findings of the effectiveness of similar models in their setting. We found that, in our setting, Prob and Good models performed worse than assigning a uniform weight to all words. We note, however, that models from Hong and Nenkova (2014) are not strictly comparable, since their word probability models were calculated after stopword exclusion, and their model that inspired our Good model was defined somewhat differently and validated using content words only. The definition of our Position model and its use in the essay scoring function S (equation 2) correspond to Hong and Nenkova (2014) average first location model for scoring summaries. Differently from their findings, this model is not effective for single words in our setting. Position models over n-grams with $n > 1$ are effective, but their prediction is in the *opposite* direction of that found for the news data – the more important materials tend to appear *later* in the lecture, as indicated by the positive r between average first position and essay score. These findings underscore the importance of paying attention to the genre of the source material when developing summarization systems.

Our summarization task incorporates elements of contrastive opinion summarization (Paul et al., 2010; Kim and Zhai, 2009), since the lecture and the reading sometimes interpret the same facts in a positive or negative light (for example, the fact that chemicals are used in fish farms is negative if compared to wild fish, but not so if compared to other farm-raised foods like poultry). Relationships between aspect and sentiment (Brody and Elhadad, 2010; Lazaridou et al., 2013) are also relevant, since aspects of the same fact are emphasized with different evaluations (the quantity vs the variety of species that go into fish meal for farmed fish). We hypothesize that units participating in sentiment and aspect contrasts are of higher importance; this is a direction for future work.

8 Conclusion

In this paper, we addressed the task of automatically assigning importance scores to parts of a lecture that is to be summarized as part of an English language proficiency test. We investigated the optimal units of information to which importance should be assigned, as well as a variety of importance scoring models, drawing on the news summarization and essay scoring literature.

We found that bigrams and trigrams were generally more effective than unigrams and 4-grams across importance models, with some exceptions.

We also found that the most effective importance models are those that equate importance of an n-gram with its preferential use in higher-scoring essays than in lower-scoring ones, above and beyond merely looking at the n-grams used in good essays. This demonstrates the utility of using not only gold, high-quality human summaries, but also sub-standard ones when developing content importance models.

Additional importance criteria that are intrinsic to the lecture, as well as those that capture contrast with a different source discussing the same topic, were also found to be reasonably effective. Since different importance models often select different items as most important, we intend to investigate complementarity of the different models.

Finally, our results highlight that the effectiveness of an importance model depends on the genre of the source text. Thus, while a first sentence baseline is very competitive in news summarization, we found that important information tends *not* to be located in the opening sentences in our data (these tend to provide general, introductory information), but appears later on, when more detailed, specific claims are put forward.

References

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

Yigal Attali. 2011. A Differential Word Use Measure for Content Analysis in Automated Essay Scoring. *ETS Research Report*, RR-11-36.

Samuel Brody and Noemie Elhadad. 2010. An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 804–812, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jill Burstein, Michael Flor, Joel Tetreault, Nitin Madnani, and Steven Holtzman. 2012. Examining Linguistic Characteristics of Paraphrase in Test-Taker Summaries. *ETS Research Report*, RR-12-18.

Keelan Evanini, Shasha Xie, and Klaus Zechner. 2013. Prompt-based content scoring for automated spoken language assessment. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 157–162, Atlanta, Georgia, June. Association for Computational Linguistics.

Peter Foltz, Lynn Streeter, Karen Lochbaum, and Thomas Landauer. 2013. Implementation and Application of the Intelligent Essay Assessor. In Mark Shermis and Jill Burstein, editors, *Handbook of automated essay evaluation: Current applications and new directions*, pages 68–88. New York: Routhledge.

Olga Gurevich and Paul Deane. 2007. Document similarity measures to distinguish native vs. non-native essay writers. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 49–52, Rochester, New York, April. Association for Computational Linguistics.

Kai Hong and Ani Nenkova. 2014. Improving the estimation of word importance for news multi-document summarization. In *The Conference of the European Chapter of the Association for Computational Linguistics*, Gottenberg, Sweden, April. Association for Computational Linguistics.

Hyun Duk Kim and ChengXiang Zhai. 2009. Generating comparative summaries of contradictory opinions in text. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 385–394, New York, NY, USA. ACM.

Angeliki Lazaridou, Ivan Titov, and Caroline Sporleder. 2013. A bayesian model for joint unsupervised induction of sentiment, aspect and discourse representations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1630–1639, Sofia, Bulgaria, August. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of ACL workshop: Text summarization branches out*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Quinn McNemar. 1955. *Psychological Statistics*. New York: J. Wiley and Sons, 2nd edition.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Human Language Technologies 2004: The Conference of the North American Chapter of the Association for Computational Linguistics*, pages 145–152, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.

Michael J. Paul, ChengXiang Zhai, and Roxana Girju. 2010. Summarizing contrastive viewpoints in opinionated text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 66–76, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lia Plakans and Atta Gebril. 2013. Using multiple texts in an integrated writing assessment: Source text use as a predictor of score. *Journal of Second Language Writing*, 22:217–230.

Mark Shermis and Jill Burstein, editors. 2013. *Handbook of Automated Essay Evaluation: Current Applications and Future Directions*. New York: Routledge.