Automatically Extracting Topical Components for a Response-to-Text Writing Assessment

Zahra Rahimi and **Diane Litman**
Intelligent Systems Program & Learning Research and Development Center
University of Pittsburgh
Pittsburgh, PA 15260
{zar10,dlitman}@pitt.edu

Abstract

We investigate automatically extracting multi-word topical components to replace information currently provided by experts that is used to score the Evidence dimension of a writing in response to text assessment. Our goal is to reduce the amount of expert effort and improve the scalability of an automatic scoring system. Experimental results show that scoring performance using automatically extracted data-driven topical components is promising.

1 Introduction

Automatic essay scoring has increasingly been investigated in recent years. One important aspect of writing assessment, specifically in source-based writing, is evaluation of content. Different methods have been used to assess the content of essays, e.g., bag of words (Mayfield and Rose, 2013), semantic similarity (Foltz et al., 1999; Kakkonen et al., 2005; Lemaire and Dessus, 2001), content vector analysis and cosine similarity (Louis and Higgins, 2010; Higgins et al., 2006; Attali and Burstein, 2006), and Latent Dirichlet Allocation (LDA) topic modeling (Persing and Ng, 2014).

These prior studies differ from our research in several ways. Much of the prior work does not target source-based writing and thus does not make use of source materials. Approaches that do make use of source materials are typically designed to detect only if an essay is on-topic. Our source-based assessment, in contrast, is also concerned with localizing in the student essay pieces of evidence that students provided from the source material. This is because our goal is to not only score an essay, but also to provide feedback based on detailed essay content.

Various kinds of source-based assessments of content (both in essay and short answering scoring) typically require some expert work in advance. Experts have provided reference answers (Nielsen et al., 2009; Mohler et al., 2011) or manually crafted patterns (Sukkarieh et al., 2004; Makatchev and VanLahn, 2007; Nielsen et al., 2009). Using manually provided information helps increase the accuracy of a scoring system and its ability to provide meaningful feedback related to the scoring rubric. But involving experts in the scoring process is a drawback for automatically scoring at scale.

Research to reduce expert effort has been underway to increase the scalability of scoring systems. A semi-supervised method is used to reduce the amount of required hand-annotated data (Zesch et al., 2015). Text templates or patterns are automatically identified for short answer scoring (Ramachandran et al., 2015). Content importance models (Beigman Klebanov et al., 2014) are used to predict source material that students should select.

In this paper, our goal is to use natural language processing to automatically extract from source material a comprehensive list of topics which include: a) important topic words, and b) specific expressions (N-grams with $N > 1$) that students need to provide in their essays. We call this comprehensive list *"topical components"*. Automatic extraction of topical components helps to reduce expert effort before the automatic assessment process. We evaluate the usefulness of our method for extracting topical components on the Response-to-Text Assessment (RTA)

Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, pages 277–282,
San Diego, California, June 16, 2016. ©2016 Association for Computational Linguistics

Excerpt from the article: Many kids in Sauri did not attend school because their parents could not afford school fees. Some kids are needed to help with chores, such as fetching water and wood. In 2004, the schools had minimal supplies like books, paper and pencils, but the students wanted to learn. All of them worked hard with the few supplies they had. It was hard for them to concentrate, though, as there was no midday meal. By the end of the day, kids didn't have any energy.
Prompt: The author provided one specific example of how the quality of life can be improved by the Millennium Villages Project in Sauri, Kenya. Based on the article, did the author provide a convincing argument that winning the fight against poverty is achievable in our lifetime? Explain why or why not with 3-4 examples from the text to support your answer.
Essay with score of 4 on Evidence dimension: *I was convinced that winning the fight of poverty is achievable in our lifetime. Many people couldn't afford medicine or bed nets to be treated for malaria. Many children had died from this dieseuse even though it could be treated easily. But now, bed nets are used in every sleeping site. And the medicine is free of charge. Another example is that the farmers' crops are dying because they could not afford the nessacary fertilizer and irrigation. But they are now, making progess. Farmers now have fertilizer and water to give to the crops. Also with seeds and the proper tools. Third, kids in Sauri were not well educated. Many families couldn't afford school. Even at school there was no lunch. Students were exhausted from each day of school. Now, school is free. Children excited to learn now can and they do have midday meals. Finally, Sauri is making great progress. If they keep it up that city will no longer be in poverty. Then the Millennium Village project can move on to help other countries in need.*

Table 1: An excerpt from the source text, the prompt, and a high-scoring essay with highlighted evidence (Rahimi et al., 2014).

(Correnti et al., 2012; Correnti et al., 2013).

RTA is developed to assess analytical writing in response to text (Correnti et al., 2013), e.g., making claims and marshalling evidence from a source text to support a viewpoint. Automatic scoring of the Evidence dimension of the RTA was previously investigated in (Rahimi et al., 2014). The Evidence dimension evaluates how well students use selected details from a text to support and extend a key idea. A set of rubric-based features enabled by topical components manually provided by experts were used in (Rahimi et al., 2014) to automatically assess Evidence.

In this paper, we propose to use a model enabled by LDA topic modeling to automatically extract the topical components (i.e., topic words and significant N-grams ($N \geq 1$)) needed for our scoring approach[1]. We hypothesize that extracting rubric-based features based on data-driven topical components can perform as well as extracting features from manually provided topical components. Results show that our method for automatically extracting topical components is promising but still needs improvement.

2 Data

We have two datasets of student writing from two different age groups (grades 5-6 and grades 6-8) that were written in response to one prompt introduced in (Correnti et al., 2013). The student essays comprising our datasets were obtained as follows. A text was read aloud by a teacher and students followed along. The text is about a United Nations project to eradicate poverty in a rural village in Kenya. After a guided discussion of the article, students wrote an essay in response to a prompt that required them to make a claim and support it using details from the text. A small excerpt from the article, the prompt, and a sample high-scoring student essay from grades 5-6 are shown in Table 1.

Our datasets (particularly essays by students in grades 5–6) have a number of properties that may increase the difficulty of the automatic essay assessment task. For example, the essays are short and many of them are only one paragraph (the median number of paragraphs for 5–6 and 6–8 datasets are 1 and 2 respectively). Some statistics about the datasets are in Table 2.

The RTA provides rubrics along five dimensions to assess student writing, each on a scale of 1-4 (Correnti et al., 2013). In this paper we focus only on predicting the score of the Evidence dimension[2]. The essays in our datasets were scored half by experts and the rest by trained undergraduates. The corpus of grades 5–6 and 6–8 respectively consist of 1569 essays with 602 of them double-scored, and 1045 essays with all of them double-scored, for inter-rater reliability. Inter-rater agreement (Quadratic Weighted Kappa) on the double-scored portion of

[1] Unlike much LDA-enabled work, we not only make use of topic words, but also expressions clustered to a set of topics.

[2] The other RTA dimensions are Analysis, Organization, Style, and MUGS (Mechanics, Usage, Grammar, Spelling).

Dataset		Mean	SD
5–6 Grades	words	161.25	92.24
	unique words	93.27	40.57
	sentences	9.01	6.39
	paragraphs	2.04	1.83
6–8 Grades	words	218.90	111.08
	unique words	109.34	41.59
	sentences	11.98	7.17
	paragraphs	2.56	1.72

Table 2: The two dataset's statistics.

Dataset	1	2	3	4	Total
5–6 Grades	471 (30%)	594 (38%)	334 (21%)	170 (11%)	1569
6–8 Grades	250 (24%)	434 (42%)	229 (22%)	132 (13%)	1045

Table 3: Distribution of the Evidence scores.

the grades 5-6 and 6-8 corpora respectively are 0.67 and 0.73 for the Evidence dimension. The distribution of Evidence scores is shown in Table 3.

3 Extracting Topical Components

One way of obtaining topical components is to have experts manually create them using their knowledge about the text (Rahimi et al., 2014). An example subset of the components, provided by experts and used to extract the features mentioned in Section 4.2, are in Table 4. The excerpt from the text from which the "school" topic is extracted is shown in Table 1.

In this paper, we instead automatically extract the topical components. Our proposed method has 3 main steps: (1) using topic modeling to extract topics and probabilistic distribution of words, (2) using Turbo-Topic to get the significant N-grams per-topic, and (3) post-processing the Turbo-Topic output to get the topical-components.

The first step uses LDA topic modeling (Blei et al., 2003) which is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. The output of the LDA algorithm is a list of topics. Each topic is a probability distribution over words in a vocabulary.

The second step feeds the posterior distribution output of LDA over words as an input to Turbo-Topic (Blei and Lafferty, 2009) to extract significant N-grams per-topic. In Turbo-Topic, the posterior distribution output of LDA is used to annotate each word occurrence in the corpus with its most probable topic. It uses a back-off language model defined for arbitrary length expressions and a statistical co-occurrence analysis is carried out recursively to extract the most significant multi-word expressions for each topic. Finally, the resulting expressions are combined with the unigram list. One advantage of Turbo-Topic is the ability of finding significant phrases without the necessity of all words in the phrase being assigned to the topic by using the information of repeated context in the language model. For example, the N-gram "schools now serve lunch" can be distinguished as a significant N-gram for the topic "School" using the language model even if only the words "schools" and "lunch" are assigned to the topic "school" by LDA.

The third step uses the output of Turbo-Topic, which is a list of significant N-grams ($N \geq 1$) with their counts per-topic, to extract the topical components. To make different topics unique and more distinguishable, we decided to include each N-gram in only one topic. For this purpose, we use the count of N-grams in topics and assign each N-gram to the topic in which it has the highest count. The next issue is to remove the redundant information. If A and B are two N-grams in a topic and A is a subset of B, we remove the N-gram A. After processing the output of Turbo-Topic, we divide it to a list of highly important words and a list of expressions per-topic. We use a cut-off threshold and only include the top N-grams based on their counts in each topic.

4 Experiments

We configure experiments to test the validity of the hypothesis that scoring models that extract features based on automatically extracted LDA-enabled topical components can perform as well as models which extract features from topical components manually provided by experts.

4.1 Experimental Tools and Methods

All experiments use 10 fold cross validation with Random Forest as a classifier (max-depth=5). We report performance using Quadratic Weighted Kappa, a standard evaluation measure for essay assessment. Paired student t-test with p-value < 0.05 is used to measure statistical significance.

	Topic:Hospitals	Topic:Schools	Topic:Progress
a) Topic Words	care, health, hospital, doctor, disease	school, supply, fee, student, lunch	progress, four, serve, attendance, maintain
b) N-grams ($N > 1$)	Yala sub district hospital no running water electricity not medicine treatment could afford no doctor only clinical officer three kids bed two adults	kids not attend go school not afford school fees no midday meal lunch schools minimal supplies concentrate not energy	progress made just four years water connected hospital bed nets used every sleeping site kids go school now now serves lunch

Table 4: A sub-list of manually extracted a) topic words and b) specific expressions for three sample topics. They are manually provided by experts in (Rahimi et al., 2014). Some of the stop-words might have been removed from the expressions by experts.

	Topic: Hospitals	Topic: Schools	Topic: Progress
a) Topic Words	author, fight, hospital, yala, sub, 2015	school, water, food, malaria, children, free	sauri, progress, made, student, project, better
b) N-grams ($N > 1$)	common diseases win the fight against poverty also has a generator for district hospital rate is way up yala subdistrict hospital	school supplies school fees and afford it food supply midday meal paper and	made amazing progress in just four years lunch for the students school now serves water is connected to the just 4 years progress in just 4

Table 5: A sub-list of automatically extracted a) topic words and b) specific expressions for three sample topics. They are automatically extracted by the data-driven LDA-enabled model (see Section 3).

We compare results for models that extracted features from topical components with a baseline model which uses the top 500 unigrams as features (chosen based on a chi-squared feature selection method), and with an upper-bound model which is the best model reported in (Rahimi et al., 2014). The only difference between our model and the upper-bound model is that in our model the topical components were extracted automatically instead of manually.

To train LDA, we use a set of 591 not-scored essays (which are not used in our cross validation experiments) from grades 6-8, and the text. We use the LDA-C implementation (Blei et al., 2003) with default values for the parameters and seeded initialization of topics to a distribution smoothed from a randomly chosen document. The number of topics is chosen equal to the number of topics provided by experts ($K = 8$). The Turbo-Topic parameters are set as P-value = 0.001 and min-count = 10 based on our intuition that it is better to discard less. The cutoff threshold for removing less frequent N-grams is intuitively set to the top 20 most frequent N-grams in a topic.

4.2 Features

We use the same set of primarily rubric-based features introduced in (Rahimi et al., 2014) to score the Evidence dimension of RTA:

Number of Pieces of Evidence (NPE): based on the list of important words for each main topic.

Concentration (CON): a binary feature which indicates if an essay has a high concentration, defined as fewer than 3 sentences with topic words.

Specificity (SPC): a vector of integer values. Each value shows the number of examples from the text mentioned in the essay for a single topic.

Word Count (WOC): number of words.

We need the list of important words per topic to calculate the NPE and CON features, and the list of important expressions per topic to calculate SPC.

5 Results and Discussion

Sample extracted topical components are in Table 5. The shown topic labels (e.g. “Hospitals”) were assigned manually by looking at the N-grams and are only for the purpose of better understanding the output. Qualitatively comparing the extracted topical components (Table 5) with the ones provided by experts (Table 4) suggests that the method presented in Section 3 can: (1) distinguish a lot of important N-grams that students were expected to cover in their essays as pieces of evidence, and (2) group related N-grams to topics. In fact, we were able to intuitively map our learned topics to 4 of the 8 manually-produced topics; 3 of these 4 mappings are shown in Table 5. However, while some of the automatically extracted topics are of a promising quality, there is still much room for improvement.

	Model	(5-6) [n=1569]	(6-8) [n=1045]
1	Unigram baseline	0.52	0.49
2	Unigram + WOC	0.53	0.52
3	Automatic (proposed)	0.56(1)	0.53(1)
4	Automatic (proposed) minus WOC	0.54	0.51
5	Manual (upper bound)	**0.62**	**0.60**

Table 6: Performance of models using automatically extracted topical components, baseline models, and the upper-bound. Bold shows that the model significantly outperforms all other models. The numbers in parentheses show model numbers that the current model significantly outperforms.

We can think of several reasons for not being able to map all automatically extracted topics to the manually produced topics. First, the manually provided topics are based on an expert's knowledge of the text. Experts may expect some details in student essays and include these in the topic list, but students are not always able to distinguish these details to cover in their essays. In other words, the LDA-enabled model is data-driven while expert knowledge is not. If some details are not covered in our training dataset, the data-driven model is not able to distinguish them. Second, experts are able to distinguish topics and their important examples even from only a few sentences in the text. But, if topics and examples are covered in the essays by only a phrase or a few sentences, the data-driven model is not able to distinguish them as distinct topics. They will not be distinguished or will be included in other topics by our model. We also observed that some examples provided by experts are broken down to more than one N-gram in our model. For example, "less than 1 dollar a day" is broken down to two N-grams: "less than" and "1 dollar a day".

Table 6 presents the quantitative performance of our proposed model, where features for predicting RTA Evidence scores are derived using the automatically extracted topical components. The results on both datasets show that the proposed model (Model 3) significantly outperforms the unigram baseline (Model 1). However, the upper-bound model performs significantly better than all other models. There is no significant difference between the rest of the models. To better understand the role of word count (which is not impacted by topical component extraction) in Model 3, we also created Models 2 and 4. Comparing Models 1 and 4, as well as Models 2 and 3, shows that the proposed model still outperforms unigrams after matching for use of word count or not. Although the improvement is no longer significant, unigrams are less useful than our rubric-based features for providing feedback. We also note that absolute performance is lower on the grade 6–8 dataset for all models, which could be due to the larger size of the 5–6 dataset. In sum, our quantitative results indicate that rubric-based Evidence scoring without involvement of experts is promising, yielding scoring models that maintain reliability while improving validity compared to unigrams. However, the gap with the upper bound shows that our topic extraction method still needs improvement.

6 Conclusion and Future Work

We developed a natural language processing technique to automatically extract topical components (topics and significant words and expressions per-topic) relevant to a source text, as our previous approach required these to be manually defined by experts. To evaluate our method, we predicted the score for the Evidence dimension of an analytical writing in response to text assessment (RTA) for upper elementary school students. Experiments comparing the predictive utility of features based on automatically extracted topical components versus manually defined components indicated promising performance for the LDA-enabled extracted topical components. Replacing experts' work with our LDA-enabled method has the potential to better scale rubric-based Evidence scoring.

There are several areas for improvement. We need to tune all parameters. We plan to examine using supervised LDA to make use of scores, or seeded LDA where a few words for each topic are provided. We should study how the size, score distribution, and spelling errors in training data impact topical extraction and scoring. We plan to examine generality by using other RTA articles and prompts. Finally, motivated by short-answer scoring (Sakaguchi et al., 2015), we would like to integrate features needing expert resources with other (valid) features.

Acknowledgments

This research was funded by the Learning Research and Development Center. We thank R. Correnti, L. C. Matsumara and E. Wang for providing expert topical components and datasets, and H. Hashemi and the ITSPOKE group for helpful feedback.

References

Y. Attali and J. 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.

David M Blei and John D Lafferty. 2009. Visualizing topics with multi-word expressions. *arXiv preprint arXiv:0907.1013*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Richard Correnti, Lindsay Clare Matsumura, Laura S Hamilton, and Elaine Wang. 2012. Combining multiple measures of students' opportunities to develop analytic, text-based writing skills. *Educational Assessment*, 17(2-3):132–161.

R. Correnti, L.C. Matsumura, L.H. Hamilton, and E. Wang. 2013. Assessing students' skills at writing in response to texts. *Elementary School Journal*, 114(2):142–177.

Peter W Foltz, Darrell Laham, and Thomas K Landauer. 1999. The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, 1(2):939–944.

Derrick Higgins, Jill Burstein, and Yigal Attali. 2006. Identifying off-topic student essays without topic-specific training data. *Natural Language Engineering*, 12(02):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.

Benoit Lemaire and Philippe Dessus. 2001. A system to assess the semantic content of student essays. *Journal of Educational Computing Research*, 24(3):305–320.

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.

Maxim Makatchev and Kurt VanLahn. 2007. Combining bayesian networks and formal reasoning for semantic classification of student utterances.

E. Mayfield and C. Rose. 2013. Lightside: Open source machine learning for text. In M. D. Shermis and J. Burstein, editors, *A Handbook of Automated Essay Evaluation: Current Applications and New Directions*, pages 124–135.

Michael Mohler, Razvan Bunescu, and Rada Mihalcea. 2011. Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 752–762.

Rodney D Nielsen, Wayne Ward, and James H Martin. 2009. Recognizing entailment in intelligent tutoring systems. *Natural Language Engineering*, 15(04):479–501.

Isaac Persing and Vincent Ng. 2014. Modeling prompt adherence in student essays. In *ACL (1)*, pages 1534–1543.

Zahra Rahimi, Diane J Litman, Richard Correnti, Lindsay Clare Matsumura, Elaine Wang, and Zahid Kisa. 2014. Automatic scoring of an analytical response-to-text assessment. In *Intelligent Tutoring Systems*, pages 601–610. Springer.

Lakshmi Ramachandran, Jian Cheng, and Peter Foltz. 2015. Identifying patterns for short answer scoring using graph-based lexico-semantic text matching. In *Proceedings of the 10th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 97–106.

Keisuke Sakaguchi, Michael Heilman, and Nitin Madnani. 2015. Effective feature integration for automated short answer scoring. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1049–1054, Denver, Colorado, May–June.

Jana Z Sukkarieh, Stephen G Pulman, and Nicholas Raikes. 2004. Auto-marking 2: An update on the ucles-oxford university research into using computational linguistics to score short, free text responses. *International Association of Educational Assessment*.

Torsten Zesch, Michael Heilman, and Aoife Cahill. 2015. Reducing annotation efforts in supervised short answer scoring. In *Proceedings of the 10th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 124–132.