Generating Reference Texts for Short Answer Scoring Using Graph-based Summarization

Lakshmi Ramachandran[1] and **Peter Foltz**[1,2]
[1]Pearson, [2]University of Colorado
{lakshmi.ramachandran,peter.foltz}@pearson.com

Abstract

Automated scoring of short answers often involves matching a students response against one or more sample reference texts. Each reference text provided contains very specific instances of correct responses and may not cover the variety of possibly correct responses. Finding or hand-creating additional references can be very time consuming and expensive. In order to overcome this problem we propose a technique to generate alternative reference texts by summarizing the content of top-scoring student responses. We use a graph-based cohesion technique that extracts the most representative answers from among the top-scorers. We also use a state-of-the-art extractive summarization tool called MEAD. The extracted set of responses may be used as alternative reference texts to score student responses. We evaluate this approach on short answer data from Semeval 2013's Joint Student Response Analysis task.

1 Introduction

Short answer scoring is a critical task in the field of automated student assessment. Short answers contain brief responses restricted to specific terms or concepts. There is a great demand for new techniques to handle large-scale development of short-answer scoring engines. For example an individual state assessment may involve building scoring algorithms for over two hundred prompts (or questions). The past few years have seen a growth in the amount of research involved in developing better features and scoring models that would help improve short answer scoring (Higgins et al., 2014; Leacock and Chodorow, 2003). The Automated Student Assessment Prize (ASAP-SAS (2012)) competition had a short answer scoring component.

Table 1: Question text, sample reference and some top-scoring answers from a prompt in the ASAP-SAS (2012) competition.

Prompt question: "Explain how pandas in China are similar to koalas in Australia and how they both are different from pythons. Support your response with information from the article."
Sample reference answer: "Specialists are limited geographically to the area of their exclusive food source. Pythons are different in both diet or eating habits and habitat from koalas. Generalists are favored over specialists. Adaptability to change. Koalas and pandas are herbivores and pythons are carnivores."
Some top-scoring student responses: "A panda and a koala are both vegetarians. Pandas eat bamboo, and koalas eat eucalyptus leaves. Pythons are not vegetarians they eat meat, and they kill there pray by strangling them or putting venom into them." "Pandas and koalas are both endangered animals. They can only be found in certain places where their food supply is. They are different from pythons because they move to a new environment and adapt as well. They be at a loss of food and climate change."

Short answer datasets are typically provided with one or more sample human references, which are representative of ideal responses. Student responses that have a high text overlap with these human references are likely to get a higher score than those that have a poor overlap. However often these sample human references are not representative of all possible correct responses. For instance consider the question, sample reference and a set of top-scoring student responses for a prompt from the ASAP-SAS (2012) competition in Table 1. The human reference provided does not encompass all possible alternative ways of expressing the correct response.

A number of approaches have been used to extract regular expressions and score student responses. Pulman and Sukkarieh (2005) use hand-crafted patterns to capture different ways of expressing the correct answer. Bachman et al. (2002) extract tags from a model answer, which are matched with stu-

Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, 2015, pages 207–212,
Denver, Colorado, June 4, 2015. ©2015 Association for Computational Linguistics

dent responses to determine their scores. Mitchell et al. (2003) use a mark scheme consisting of a set of acceptable or unacceptable answers. This marking scheme is similar to a sample reference. Each student response is matched with these marking schemes and scored accordingly. The winner of the ASAP competition spent a lot of time and effort hand-coding regular expressions from the human samples provided, in order to obtain better matches between student responses and references (Tandalla, 2012). Although hand-crafting features might seem feasible for a few prompts, it is not an efficient technique when scoring large datasets consisting of thousands of prompts. Hence there is a need to develop automated ways of generating alternate references that are more representative of top-scoring student responses.

We use two summarization techniques to identify alternative references from top-scoring student responses for a prompt. Klebanov et al. (2014) use summarization to generate content importance models from student essays. We propose a graph-based cohesion technique, which uses text structure and semantics to extract representative responses. We also use a state-of-the-art summarization technique called MEAD (Radev et al., 2004), which extracts a summary from a collection of top-scoring responses. The novelty of our work lies in the utilization of summarization to the task of identifying suitable references to improve short-answer scoring.

2 Approach

Top-scoring responses from each prompt or question are summarized to identify alternate reference texts with which student responses could be compared to improve scoring models.

2.1 Graph-based Cohesion Technique

We use an agglomerative clustering technique to group lexico-semantically close responses into clusters or topics. The most representative responses are extracted from each of the clusters to form the set of alternate references. Just as in a cohesion-based method only the most well-connected vertices are taken to form the summary (Barzilay and Elhadad, 1997), likewise in our approach responses with the highest similarities within each cluster are selected as representatives.

Steps involved in generating summaries are:

Generating Word-Order Graphs: Each top-scoring response is first represented as a word-order graph. We use a word-order graph representation because it captures structural information in texts. Graph matching makes use of the ordering of words and context information to help identify lexical changes. According to Makatchev and VanLehn (2007) responses classified by human experts into a particular semantic class may be syntactically different. Thus word-order graphs are useful to identify representatives from a set of responses that are similar in meaning but may be structurally different.

During graph generation, each response is tagged with parts-of-speech (POS) using the Stanford POS tagger (Toutanova et al., 2003). Contiguous subject components such as nouns, prepositions are grouped to form a subject vertex, while contiguous verbs or modals are grouped into a verb vertex and so on for the other POS types. Ordering is maintained with the edges capturing subject—verb, verb—object, subject—adjective or verb—adverb type of information. Graph generation has been explained in detail in Ramachandran and Gehringer (2012).

Calculating Similarity: In this step similarities between all pairs of top-scoring responses are calculated. Similarities between pairs of responses are used to cluster them and then identify representative responses from each cluster. *Similarity* is the average of the best vertex and edge matches.

$$\begin{aligned}\mathit{Similarity}(A,B) = \; & \tfrac{1}{2}\Big(\tfrac{1}{|V_A|+|V_B|}\Big(\sum_{\forall V_A}\underset{\forall V_B}{\operatorname{argmax}}\{\mathit{sem}(V_A,V_B)\} \\ & + \sum_{\forall V_B}\underset{\forall V_A}{\operatorname{argmax}}\{\mathit{sem}(V_B,V_A)\}\Big)+ \\ & \tfrac{1}{|E_A|+|E_B|}\Big(\sum_{\forall E_A}\underset{\forall E_B}{\operatorname{argmax}}\{\mathit{sem}_e(E_A,E_B)\} \\ & + \sum_{\forall E_B}\underset{\forall E_A}{\operatorname{argmax}}\{\mathit{sem}_e(E_B,E_A)\}\Big)\Big)\end{aligned} \tag{1}$$

In equation 1 V_A and V_B are the vertices and E_A and E_B are the edges of responses A and B respectively. We identify the best semantic match for every vertex or edge in response A with a vertex or edge in response B respectively (and vice-versa). *sem* is identified using WordNet (Fellbaum, 1998).

Clustering Responses: We use an agglomerative clustering technique to group responses into clusters. The clustering algorithm starts with assigning every response in the text to its own cluster. Initially every cluster's similarity is set to 0. A cluster's similarity is the average of the similarity between all pairs of responses it contains.

We rank response pairs based on their similarity (highest to lowest) using merge sort, and assign one response in a pair to the other's cluster provided it satisfies the condition in Equation 2. The condition ensures that a response (S) that is added to a cluster (C) has high similarity, i.e., is close in meaning and context to that cluster's responses (S_C).

$$\left(C.\mathit{clusterSimilarity} - \sum_{\forall S_C \in C} \frac{\mathit{Similarity}(S, S_C)}{|C|} \right) \leq \alpha \quad (2)$$

The choice of cluster to which a response is added depends on the cluster's similarity, i.e., a response is added to the cluster with higher similarity. If both responses (in the pair) have same cluster similarities, then the larger cluster is chosen as the target. If cluster similarity and the number of responses are the same, then the target is selected randomly.

Identifying Representatives: In this step the most representative responses from each cluster are identified. The aim is to identify the smallest set of representatives that *cover* every other response in the cluster. We use a list heuristic to handle this problem (Avis and Imamura, 2007). We order responses in every cluster based on (a) decreasing order of their average similarity values, and (b) decreasing order of the number of responses they are adjacent to.

Our approach ensures that responses with the highest semantic similarity that cover previously uncovered responses are selected. Representatives from all clusters are grouped together to generate the representative responses for a prompt.

2.2 MEAD

We use MEAD as an alternative summarization approach. Radev et al. (2004) proposed the use an automated multi-document summarization technique called MEAD. MEAD was developed at the University of Michigan as a centroid-based summarization approach. MEAD is an extractive summarization approach that relies on three features: position, centroid and the length of sentences to identify the summary. MEAD's classifier computes a score for each sentence in the document using a linear combination of these three features. Sentences are then ranked based on their scores and the top ranking sentences are extracted to generate summaries. The extraction can be restricted to the top N words to generate a summary of specified length.

In our study each document contains a list of top-scoring responses from the dataset, i.e., each top-scoring response would constitute a sentence. For our study we use MEAD[1] to extract summaries of length that match the lengths of the summaries generated by the graph-based cohesion technique.

3 Experiment

3.1 Data

Semeval's Student Response Analysis (SRA) corpus contains short answers from two different sources: Beetle and SciEntsBank (Dzikovska et al., 2013)[2]. Beetle contains responses extracted from transcripts of interactions between students and the Beetle II tutoring system (Dzikovska et al., 2010). The SciEntsBank dataset contains short responses to questions collected by Nielsen et al. (2008).

Beetle contains 47 questions and 4380 student responses, and SciEntsBank contains 135 questions and 5509 student responses (Dzikovska et al., 2013). Each dataset is classified as: (1) 5-way, (2) 3-way and (3) 2-way. The data in the SRA corpus was annotated as follows for the 5-way classification: *correct:* student response that is correct, *partially_correct_incomplete:* response that is correct but does not contain all the information in the reference text, *contradictory:* response that contradicts the reference answer, *irrelevant:* response that is relevant to the domain but does not contain information in the reference, *non_domain:* response is not relevant to the domain. The 3-way classification contains the contradictory, correct and incorrect classes, while the 2-way classification contains correct and incorrect classes.

Dzikovska et al. (2013) provide a summary of the results achieved by teams that participated in this task. Apart from the dataset, the organizing committee also released code for a baseline, which included lexical overlap measures. These measures

[1] We use the code for MEAD (version 3.10) available at http://www.summarization.com/mead/.

[2] The data is available at http://www.cs.york.ac.uk/semeval-2013/task7/index.php?id=data

Table 2: Comparing performance of system-generated summaries of top-scoring short answers with the performance of sample reference texts provided for the Semeval dataset.

		5-way		3-way		2-way	
Data Type	**System**	**F1-overall**	**Weighted-F1**	**F1-overall**	**Weighted-F1**	**F1-overall**	**Weighted-F1**
Beetle	Baseline features (Dzikovska et al., 2013)	0.424	0.483	0.552	0.578	0.788	
	Graph (∼62 words)	0.436	0.533	**0.564**	**0.587**	**0.794**	**0.803**
	MEAD (∼63 words)	**0.446**	**0.535**	0.537	0.558	0.744	0.757
SciEntsBank	Baseline features (Dzikovska et al., 2013)	0.375	0.435	0.405	0.523	0.617	
	Graph (∼39 words)	0.372	0.458	**0.438**	**0.567**	**0.644**	**0.658**
	MEAD (∼40 words)	**0.379**	**0.461**	0.429	0.554	0.631	0.645

Table 3: Comparing f-measures (f) and mean cosines (cos) of every class for features generated by graph and MEAD summaries.

		5-way					3-way			2-way	
Classes	**Feature**	**correct**	**partially_correct_incomplete**	**contradictory**	**non_domain**	**irrelevant**	**correct**	**contradictory**	**incorrect**	**correct**	**incorrect**
Beetle											
MEAD	f	0.702	**0.443**	**0.416**	**0.667**	0.000	0.687	0.400	0.523	0.679	0.809
Graph	f	**0.736**	0.400	0.404	0.640	0.000	**0.732**	**0.422**	**0.539**	0.747	**0.840**
MEAD	cos	0.690	0.464	**0.438**	0.058	**0.319**	0.690	**0.438**	0.387	0.690	**0.408**
Graph	cos	**0.720**	**0.470**	0.425	**0.065**	0.286	**0.720**	0.425	**0.388**	**0.720**	0.404
SciEntsBank											
MEAD	f	0.601	**0.332**	0.082	NA	**0.500**	0.563	**0.062**	**0.661**	0.528	0.733
Graph	f	**0.617**	0.302	0.087	NA	0.482	**0.605**	0.059	0.649	**0.548**	**0.741**
MEAD	cos	0.441	0.337	0.337	0.138	0.268	0.441	0.337	0.298	0.441	0.305
Graph	cos	**0.498**	**0.372**	**0.350**	**0.229**	**0.271**	**0.498**	**0.350**	**0.316**	**0.498**	**0.323**

compute the degree of overlap between student responses and sample reference texts and the prompt or question texts. Both human references as well as question texts were provided with the dataset. The lexical overlap measures include: (1) Raw count of the overlaps between student responses and the sample reference and question texts, (2) Cosine similarity between the compared texts, (3) Lesk similarity, which is the sum of square of the length of phrasal overlaps between pairs of texts, normalized by their lengths (Pedersen et al., 2002) and (4) f-measure of the overlaps between the compared texts[3]. These four features are computed for the sample reference text and the question text, resulting in a total of eight features. We compute these eight features for every system and compare their raw and weighted (by their class distributions) f-measure values.

3.2 Results and Discussion

The graph-based cohesion technique produced summaries containing an average of 62 words for Beetle and an average of 39 words for SciEntsBank. Therefore, we chose to extract summaries containing nearly the same number of words using the MEAD summarization tool.

From the results in Table 2[4] we see that, compared to the baseline approach, the summarization approaches are better at scoring short answers. We also tested the use of all top-scoring student responses as alternate references (i.e. with no summarization). These models perform worse than the baseline, producing an average *decrease* in overall f-measure of 14.7% for Beetle and 14.3% for SciEntsBank. This suggests the need for a summarization technique. Our results indicate that the summarizers produce representative sentences that are more useful for scoring than just the sample reference text. MEAD performs better on the 5-way task while the graph-based cohesion approach performs well on 3-way and 2-way classification tasks.

In the case of both the datasets, the performance of the graph-based approach on the "correct" class is higher. We looked at the average cosine similarity for data from each class with their corre-

[3] f-measure is the harmonic mean of the precision and recall of the degree of overlaps between two texts. Precision is computed as the number of overlaps divided by the length of student response, while recall of overlap is computed as the degree of overlap divided by the number of tokens in the human reference text.

[4] We report results only on the unseen answers test set from Semeval because the train and test sets contain data from different prompts for the unseen domains and unseen questions sets. Summaries generated from the top-scoring responses from one set of prompts or questions in the train set may not be relevant to different prompts in the other test sets.

Table 4: Comparing references generated by the summarizers with a sample reference for a prompt from the Beetle dataset.

Sample Reference: "Terminal 1 and the positive terminal are separated by the gap OR Terminal 1 and the positive terminal are not connected. OR Terminal 1 is connected to the negative battery terminal. OR Terminal 1 is not separated from the negative battery terminal. OR Terminal 1 and the positive battery terminal are in different electrical states"	**Graph-based Cohesion:** "The terminal is not connected to the positive battery terminal. OR The terminals are not connected. OR The positive battery terminal and terminal 1 are not connected. OR Because there was not direct connection between the positive terminal and bulb terminal 1. OR Terminal one is connected to the negative terminal and terminal 1 is separated from the positive terminal by a gap. OR The positive battery terminal is separated by a gap from terminal 1."	**MEAD:** "Positive battery terminal is separated by a gap from terminal 1. OR Terminal 1 is not connected to the positive terminal. OR Because there was not direct connection between the positive terminal and bulb terminal 1. OR The terminals are not connected. OR Because they are not connected. OR Terminal 1 is connected to the negative battery terminal. OR The two earnt connected."

sponding reference texts (Table 3). Magnitude of the average cosine between student responses and the reference texts for classes such as non_domain and partially_correct_incomplete in Beetle and for non_domain, partially_correct_incomplete, contradictory and irrelevant in SciEntsBank are higher in case of the graph-based approach than MEAD. As a result, the graph's features tend to classify more data points as correct, leaving fewer data points to be classified into the other classes, thus producing lower f-measures in both datasets.

In the case of 3-way and 2-way classifications, performance on the correct class was higher for the graph-based approach (Table 3). The cosine similarity between the correct data and the summaries from the graph-based approach are higher than the cosines between the correct data and MEAD's summaries. The graph-based approach tends to predict more of the correct data points accurately, resulting in an improvement in the graph-based approach's performance. A similar trend was observed in the case of the 2-way classification.

Sample reference and representatives from the graph-based approach and MEAD for question BULB_C_VOLTAGE_EXPLAIN_WHY1 from Beetle are listed in Table 4. The samples follow the structure X and Y are <relation> OR X <relation> Y. A correct response such as "The terminals are not connected." would get a low match with these samples. Both the graph-based approach and MEAD extract references that may be structurally different but have the same meaning.

The team that performed best on the Semeval competition on both the Beetle and SciEntsBank datasets for the unseen answers task (Heilman and Madnani, 2013), used the baseline features (listed above) as part of their models. CoMeT was another team that performed well on Beetle on the unseen answers dataset (Ott et al., 2013). They did not use the baseline features directly but did use the sample reference text to generate several text overlap measures. Since the best performing models used sample references to generate useful features, the use of representative sentences generated by a summarization approach is likely to help boost the performance of these models. We have not been able to show the improvement to the best models from Semeval since the code for the best models have not been made available. These alternate references also generate improved baselines, thus encouraging teams participating in competitions to produce better models.

4 Conclusion

In this paper we demonstrated that an automated approach to generating alternate references can improve the performance of short answer scoring models. Models would benefit a great deal from the use of alternate references that are likely to cover more types of correct responses than the sample. We evaluated two summarization techniques on two short answer datasets: Beetle and SciEntsBank made available through the Semeval competition on student response analysis. We showed that references generated from the top-scoring responses by the graph-based cohesion approach and by MEAD performed better than the baseline containing the sample reference.

The results indicate that the approach can be successfully applied for improving scoring of short answers responses. These results have direct applications to automated tutoring systems, where students are in a dialogue with a computer-based agent and the system must match the student dialogue against a set of reference responses. In each of these cases, the technique provides a richer set of legal reference texts and it can be easily incorporated as a preprocessing step before comparisons are made to the student responses.

References

ASAP-SAS. 2012. Scoring short answer essays. ASAP short answer scoring competition system description. `http://www.kaggle.com/c/asap-sas/`.

David Avis and Tomokazu Imamura. 2007. A list heuristic for vertex cover. volume 35, pages 201–204. Elsevier.

Lyle F Bachman, Nathan Carr, Greg Kamei, Mikyung Kim, Michael J Pan, Chris Salvador, and Yasuyo Sawaki. 2002. A reliable approach to automatic assessment of short answer free responses. In *Proceedings of the 19th international conference on Computational linguistics-Volume 2*, pages 1–4. Association for Computational Linguistics.

Regina Barzilay and Michael Elhadad. 1997. Using lexical chains for text summarization. In *In Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, pages 10–17.

Myroslava O Dzikovska, Johanna D Moore, Natalie Steinhauser, Gwendolyn Campbell, Elaine Farrow, and Charles B Callaway. 2010. Beetle ii: a system for tutoring and computational linguistics experimentation. In *Proceedings of the ACL 2010 System Demonstrations*, pages 13–18. Association for Computational Linguistics.

Myroslava O Dzikovska, Rodney D Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa T Dang. 2013. Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. Technical report, DTIC Document.

Christiane Fellbaum. 1998. Wordnet: An electronic lexical database. *MIT Press*.

Michael Heilman and Nitin Madnani. 2013. ETS: Domain adaptation and stacking for short answer scoring. In *Proceedings of the 2nd joint conference on lexical and computational semantics*, volume 2.

Derrick Higgins, Chris Brew, Michael Heilman, Ramon Ziai, Lei Chen, Aoife Cahill, Michael Flor, Nitin Madnani, Joel Tetreault, Dan Blanchard, Diane Napolitano, Chong Min Lee, and John Blackmore. 2014. Is getting the right answer just about choosing the right words? The role of syntactically-informed features in short answer scoring. *arXiv*.

Beata Beigman Klebanov, Nitin Madnani, Swapna Somasundaran, and Jill Burstein. 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. Association for Computational Linguistics.

Claudia Leacock and Martin Chodorow. 2003. C-rater: Automated scoring of short-answer questions. In *Computers and the Humanities*, volume 37, pages 389–405. Springer.

Maxim Makatchev and Kurt VanLehn. 2007. Combining bayesian networks and formal reasoning for semantic classification of student utterances. In *Proceedings of the 2007 Conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work*, pages 307–314, Amsterdam, The Netherlands.

Tom Mitchell, Nicola Aldridge, and Peter Broomhead. 2003. Computerised marking of short-answer free-text responses. In *Manchester IAEA conference*.

Rodney D Nielsen, Wayne Ward, James H Martin, and Martha Palmer. 2008. Annotating students' understanding of science concepts. In *LREC*.

Niels Ott, Ramon Ziai, Michael Hahn, and Detmar Meurers. 2013. Comet: integrating different levels of linguistic modeling for meaning assessment.

Ted Pedersen, Satanjeev Banerjee, and Siddharth Patwardhan. 2002. Lesk similarity. `http://search.cpan.org/dist/Text-Similarity/lib/Text/Similarity/Overlaps.pm`.

Stephen G Pulman and Jana Z Sukkarieh. 2005. Automatic short answer marking. In *Proceedings of the second workshop on Building Educational Applications Using NLP*, pages 9–16.

Dragomir R Radev, Hongyan Jing, Małgorzata Styś, and Daniel Tam. 2004. Centroid-based summarization of multiple documents. *Information Processing & Management*, 40(6):919–938.

Lakshmi Ramachandran and Edward F. Gehringer. 2012. A word-order based graph representation for relevance identification (poster). *Proceedings of the 21st ACM Conference on Information and Knowledge Management*, pages 2327–2330, October.

Luis Tandalla. 2012. Scoring short answer essays. ASAP short answer scoring competition–Luis Tandalla's approach. `https://kaggle2.blob.core.windows.net/competitions/kaggle/2959/media/TechnicalMethodsPaper.pdf`.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *In Proceedings of HLT-NAACL*, pages 252–259.