Automatic detection of plagiarized spoken responses

Keelan Evanini and Xinhao Wang
Educational Testing Service
660 Rosedale Road, Princeton, NJ, USA
{kevanini,xwang002}@ets.org

Abstract

This paper addresses the task of automatically detecting plagiarized responses in the context of a test of spoken English proficiency for non-native speakers. A corpus of spoken responses containing plagiarized content was collected from a high-stakes assessment of English proficiency for non-native speakers, and several text-to-text similarity metrics were implemented to compare these responses to a set of materials that were identified as likely sources for the plagiarized content. Finally, a classifier was trained using these similarity metrics to predict whether a given spoken response is plagiarized or not. The classifier was evaluated on a data set containing the responses with plagiarized content and non-plagiarized control responses and achieved accuracies of 92.0% using transcriptions and 87.1% using ASR output (with a baseline accuracy of 50.0%).

1 Introduction

The automated detection of plagiarism has been widely studied in the domain of written student essays, and several online services exist for this purpose.[1] In addition, there has been a series of shared tasks using common data sets of written language to compare the performance of a variety of approaches to plagiarism detection (Potthast et al., 2013). In contrast, the automated detection of plagiarized spoken responses has received little attention from both the NLP and assessment communities, mostly due to the limited application of automated speech scoring for the types of spoken responses that could be affected by plagiarism. Due to a variety of factors, though, this is likely to change in the near future, and the automated detection of plagiarism in spoken language will become an increasingly important application.

First of all, English continues its spread as the global language of education and commerce, and there is a need to assess the communicative compentance of high volumes of highly proficient non-native speakers. In order to provide a valid evaluation of the complex linguistic skills that are necessary for these speakers, the assessment must contain test items that elicit spontaneous speech, such as the Independent and Integrated Speaking items in the TOEFL iBT test (ETS, 2012), the Retell Lecture item in the Pearson Test of English Academic (Longman, 2010), and the oral interview in the IELTS Academic assessment (Cullen et al., 2014). However, with the increased emphasis on complex linguistic skills in assessments of non-native speech, there is an increased chance that test takers will prepare canned answers using test preparation materials prior to the examination. Therefore, research should also be conducted on detecting spoken plagiarized responses in order to prevent this type of cheating strategy.

In addition, there will also likely be an increase in spoken language assessments for native speakers in the K-12 domain in the near future. Curriculum developers and assessment designers are recognizing that the assessment of spoken communication skills is important for determining a student's college readiness. For example, the Common Core State Standards include Speaking & Listening English Language Arts standards for each grade that pertain to a student's ability to communicate information and ideas using spoken language.[2] In order to assess these standards, it

[1] For example, http://turnitin.com/en_us/features/originalitycheck, http://www.grammarly.com/plagiarism-checker/, and http://www.paperrater.com/plagiarism_checker.

[2] http://www.corestandards.org/ELA-Literacy/SL/

Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications , pages 22–27,
Baltimore, Maryland USA, June 26, 2014.

will be necessary to develop standardized assessments for the K-12 domain that contain items eliciting spontaneous speech from the student, such as presentations, group discussions, etc. Again, with the introduction of these types of tasks, there is a risk that a test taker's spoken response will contain prepared material drawn from an external source, and there will be a need to automatically detect this type of plagiarism on a large scale, in order to provide fair and valid assessments.

In this paper, we present an initial study of automated plagiarism detection on spoken responses containing spontaneous non-native speech. A data set of actual plagiarized responses was collected, and text-to-text similarity metrics were applied to the task of classifying responses as plagiarized or non-plagiarized.

2 Previous Work

A wide variety of techniques have been employed in previous studies for the task of detecting plagiarized written documents, including n-gram overlap (Lyon et al., 2006), document fingerprinting (Brin et al., 1995), word frequency statistics (Shivakumar and Garcia-Molina, 1995), Information Retrieval-based metrics (Hoad and Zobel, 2003), text summarization evaluation metrics (Chen et al., 2010), WordNet-based features (Nahnsen et al., 2005), and features based on shared syntactic patterns (Uzuner et al., 2005). This task is also related to the widely studied task of paraphrase recognition, which benefits from similar types of features (Finch et al., 2005; Madnani et al., 2012). The current study adopts several of these features that are designed to be robust to the presence of word-level modifications between the source and the plagiarized text; since this study focuses on spoken responses that are reproduced from memory and subsequently processed by a speech recognizer, metrics that rely on exact matches are likely to perform sub-optimally. To our knowledge, no previous work has been reported on automatically detecting similar spoken documents, although research in the field of Spoken Document Retrieval (Haputmann, 2006) is relevant.

Due to the difficulties involved in collecting corpora of actual plagiarized material, nearly all published results of approaches to the task of plagiarism detection have relied on either simulated plagiarism (i.e., plagiarized texts generated by experimental human participants in a controlled environment) or artificial plagiarism (i.e., plagiarized texts generated by algorithmically modifying a source text) (Potthast et al., 2010). These results, however, may not reflect actual performance in a deployed setting, since the characteristics of the plagiarized material may differ from actual plagiarized responses. To overcome this limitation, the current study is based on a set of actual plagiarized responses drawn from a large-scale assessment.

3 Data

The data used in this study was drawn from the TOEFL® Internet-based test (TOEFL® iBT), a large-scale, high-stakes assessment of English for non-native speakers, which assesses English communication skills for academic purposes. The Speaking section of TOEFL iBT contains six tasks, each of which requires the test taker to provide an extended response containing spontaneous speech. Two of the tasks are referred to as Independent tasks; these tasks cover topics that are familiar to test takers and ask test takers to draw upon their own ideas, opinions, and experiences in a 45-second spoken response (ETS, 2012). Since these two Independent tasks ask questions that are not based on any stimulus materials that were provided to the test taker (such as a reading passage, figure, etc.), the test takers can provide responses that contain a wide variety of specific examples.

In some cases, test takers may attempt to game the assessment by memorizing canned material from an external source and adapting it to a question that is asked in one of the Independent tasks. This type of plagiarism can affect the validity of a test taker's speaking score; however, it is often difficult even for trained human raters to recognize plagiarized spoken responses, due to the large number and variety of external sources that are available from online test preparation sites.

In order to better understand the strategies used by test takers who incorporated material from external sources into their spoken responses and to develop a capability for automated plagiarism detection for speaking items, a data set of operational spoken responses containing potentially plagiarized material was collected. This data set contains responses that were flagged by human raters as potentially containing plagiarized material and then subsequently reviewed by rater supervisors. In the review process, the responses were transcribed and compared to external source materi-

als obtained through manual internet searches; if it was determined that the presence of plagiarized material made it impossible to provide a valid assessment of the test taker's performance on the task, the response was assigned a score of 0. This study investigates a set of 719 responses that were flagged as potentially plagiarized between October 2010 and December 2011; in this set, 239 responses were assigned a score of 0 due to the presence of a significant amount of plagiarized content from an identified source. This set of 239 responses is used in the experiments described below.

During the process of reviewing potentially plagiarized responses, the raters also collected a data set of external sources that appeared to have been used by test takers in their responses. In some cases, the test taker's spoken response was nearly identical to an identified source; in other cases, several sentences or phrases were clearly drawn from a particular source, although some modifications were apparent. Table 1 presents a sample source that was identified for several of the 239 responses in the data set.[3] Many of the plagiarized responses contained extended sequences of words that directly match idiosyncratic features of this source, such as the phrases "how romantic it can ever be" and "just relax yourself on the beach."

In total, 49 different source materials were identified for all of the potentially plagiarized responses in the corpus.[4] In addition to the source materials and the plagiarized responses, a set of non-plagiarized control responses was also obtained in order to conduct classification experiments between plagiarized and non-plagiarized responses. Since the plagiarized responses were collected over the course of more than one year, they were drawn from many different TOEFL iBT test forms; in total, the 239 plagiarized responses comprise 103 distinct Independent test questions. Therefore, it was not practical to obtain control data from all of the test items that were represented in the plagiarized set; rather, approximately 300 responses were extracted from each of the four test

[3]This source is available from several online test preparation websites, for example `http://www.mhdenglish.com/eoenglish_article_view_1195.html`.

[4]A total of 39 sources were identified for the set of 239 responses in the Plagiarized set; however, all 49 identified sources were used in the experiments in order to make the experimental design more similar to an operational set-up in which the exact set of source texts that will be represented in a given set of plagiarized responses is not known.

Well, the place I enjoy the most is a small town located in France. I like this small town because it has very charming ocean view. I mean the sky there is so blue and the beach is always full of sunshine. You know how romantic it can ever be, just relax yourself on the beach, when the sun is setting down, when the ocean breeze is blowing and the seabirds are singing. Of course I like this small French town also because there are many great French restaurants. They offer the best seafood in the world like lobsters and tuna fishes. The most important, I have been benefited a lot from this trip to France because I made friends with some gorgeous French girls. One of them even gave me a little watch as a souvenir of our friendship.

Table 1: Sample source passage used in plagiarized responses

items that were most frequently represented in the set of plagiarized responses. Table 2 provides a summary of the three data sets used in the study, along with summary statistics about the length of the responses in each set.

Data Set	N	Number of Words	
		Mean	Std. Dev.
Sources	49	122.5	36.5
Plagiarized	239	109.1	18.9
Control	1196	84.9	24.1

Table 2: Summary of the data sets

As Table 2 shows, the plagiarized responses are on average a little longer than the control responses. This is likely due to the fact that the plagiarized responses contain a large percentage of memorized material, which the test takers are able to produce using a fast rate of speech, since they had likely rehearsed the content several times before taking the assessment.

4 Methodology

The general approach taken in this study for determining whether a spoken response is plagiarized or not was to compare its content to the content of each of the source materials that had been identified for the responses in this corpus. Given a test response, a comparison was made with each

of the 49 reference sources using the following 9 text-to-text similarity metrics: 1) Word Error Rate (WER), or edit distance between the response and the source; 2) TER, similar to WER, but allowing shifts of words within the text at a low edit cost (Snover et al., 2006); 3) TER-Plus, an extension of TER that includes matching based on paraphrases, stemming, and synonym substitution (Snover et al., 2008); 4) a WordNet similarity metric based on presence in the same synset;[5] 5) a WordNet similarity metric based on the shortest path between two words in the *is-a* taxonomy; 6) a WordNet similarity metric similar to (5) that also takes into account the maximum depth of the taxonomy in which the words occur (Leacock and Chodorow, 1998); 7) a WordNet similarity metric based on the depth of the Least Common Subsumer of the two words (Wu and Palmer, 1994); 8) Latent Semantic Analysis, using a model trained on the British National Corpus (BNC, 2007); 9) BLEU (Papineni et al., 2002). Most of these similarity metrics (with the exception of WER and TER) are expected to be robust to modifications between the source text and the plagiarized response, since they do not rely on exact string matches.

Each similarity metric was used to compute 4 different features comparing the test response to each of the 49 source texts: 1) the document-level similarity between the test response and the source text; 2) the single maximum similarity value from a sentence-by-sentence comparison between the test response and the source text; 3) the average of the similarity values for all sentence-by-sentence comparisons between the test response and the source text; 4) the average of the maximum similarity values for each sentence in the test response, where the maximum similarity of a sentence is obtained by comparing it with each sentence in the source text. The intuition behind using the features that compare sentence-to-sentence similarity as opposed to only the document-level similarity feature is that test responses may contain a combination of both passages that were memorized from a source text and novel content. Depending on the amount of the response that was plagiarized, these types of responses may also receive a score of 0; so, in order to also detect these responses as plagiarized, a sentence-by-sentence comparison approach may be more effective.

The experiments described below were conducted using both human transcriptions of the spoken responses as well as the output from an automated speech recognition (ASR) system. The ASR system was trained on approximately 800 hours of TOEFL iBT responses; the system's WER on the data used in this study was 0.411 for the Plagiarized set and 0.362 for the Control set. Since the ASR output does not contain sentence boundaries, these were obtained using a Maximum Entropy sentence boundary detection system based on lexical features (Chen and Yoon, 2011). Before calculating the similarity features, all of the texts were preprocessed to normalize case, segment the text into sentences, and remove disfluencies, including filled pauses (such as *uh* and *um*) and repeated words. No stemming was performed on the words in the texts for this study.

5 Results

As described in Section 4, 36 similarity features were calculated between each spoken response and each of the 49 source texts. In order to examine the performance of these features in discriminating between plagiarized and non-plagiarized responses, classification experiments were conducted on balanced sets of Plagiarized and Control responses, and the results were averaged using 1000 random subsets of 239 responses from the Control set.[6] In addition, the following different feature sets were compared: All (all 36 features), Doc (the 9 document-level features), and Sent (the 27 features based on sentence-level comparisons). The J48 decision tree model from the Weka toolkit (with the default parameter settings) was used for classification, and 10-fold cross-validation was performed using both transcriptions and ASR output. Table 3 presents the results of these experiments, including the means (and standard deviations) of the accuracy and kappa (κ) values (for all experiments, the baseline accuracy is 50%).

6 Discussion and Future Work

As Table 3 shows, the classifier achieved a higher accuracy when using the 9 document-level similarity features compared to using the 27 sentence-

[5] For the WordNet-based similarity metrics, the similarity scores for pairs of words were combined to obtain document- and sentence-level similarity scores by taking the average maximum pairwise similarity values, similar to the sentence-level similarity feature defined in (4) below.

[6] Experiments were also conducted using the full Control set, and the results showed a similar relative performance of the feature sets.

Text	Features	Accuracy	κ
Trans.	All	0.903 (0.01)	0.807 (0.02)
	Doc	0.920 (0.01)	0.839 (0.02)
	Sent	0.847 (0.01)	0.693 (0.03)
ASR	All	0.852 (0.02)	0.703 (0.03)
	Doc	0.871 (0.01)	0.742 (0.03)
	Sent	0.735 (0.02)	0.470 (0.04)

Table 3: Mean Accuracy and κ values (and standard deviations) for classification results using the 239 responses in the Plagiarized set and 1000 random subsets of 239 responses from the Control set

level similarity features. In addition, the combined set of 36 features resulted in a slightly lower performance than when only the 9 document-level features were used. This suggests that the sentence level features are not as robust as the document-level features, probably due to the increased likelihood of chance similarities between sentences in the response and a source text. Despite the fact that the plagiarized spoken responses in this data set may contain some original content (in particular, introductory material provided by the test taker in an attempt to make the plagiarized content seem more relevant to the specific test question), it appears that the document-level features are most effective. Table 3 also indicates that the performance of the classifier decreases by approximately 5% - 10% when ASR output is used. This indicates that the similarity metrics are reasonably robust to the presence of speech recognition errors in the text, and that the approach is viable in an operational setting in which transcriptions of the spoken responses are not available.

A more detailed error analysis indicates that the precision of the classifier, with respect to the Plagiarized class, is higher than the recall: on the transcriptions, the average precision using the Doc features was 0.948 (s.d.= 0.01), whereas the average recall was 0.888 (s.d.=0.01); for the ASR set, the average precision was 0.904 (s.d.=0.02), whereas the average recall was 0.831 (s.d.=0.02). This means that the rate of false positives produced by this classifier is somewhat lower than the rate of false negatives. In an operational scenario, an automated plagiarized spoken response detection system such as this one would likely be deployed in tandem with human raters to review the results and provide a final decision about whether a given spoken response was plagiarized or not. In that case, it may be desirable to tune the classifier parameters to increase the recall so that fewer cases of plagiarism would go undetected, assuming that there are suffient human reviewers available to process the increased number of false positives that would result from this approach. Improving the classifier's recall is also important for practical applications of this approach, since the distribution of actual responses is heavily imbalanced in favor of the non-plagiarized class. The current set of experiments only used a relatively small Control set of 1196 responses for which transcriptions could be obtained in a cost effective manner in order to be able to compare the system's performance using transcriptions and ASR output. Since there was only a minor degradation in performance when ASR output was used, future experiments will be conducted using a much larger Control set in order to approximate the distribution of categories that would be observed in practice.

One drawback of the method described in this study is that it requires matching source texts in order to detect a plagiarized spoken response. This means that plagiarized spoken responses based on a given source text will not be detected by the system until the appropriate source text has been identified, thus limiting the system's recall. Besides attempting to obtain additional source texts (either manually, as was done for this study, or by automated means), this could also be addressed by comparing a test response to all previously collected spoken responses for a given population of test takers in order to flag pairs of similar responses. While this method would likely produce a high number of false positives when the ASR output was used, due to chance similarities between two responses in a large pool of test taker responses resulting from imperfect ASR, performance could be improved by considering additional information from the speech recognizer when computing the similarity metrics, such as the N-best list. Additional sources of information that could be used for detecting plagiarized responses include stylistic patterns and prosodic features; for example, spoken responses that are reproduced from memory likely contain fewer filled pauses and have a faster rate of speech than non-plagiarized responses; these types of non-lexical features should also be investigated in future research into the detection of plagiarized spoken responses.

Acknowledgments

We would like to thank Beata Beigman Klebanov, Dan Blanchard, Nitin Madnani, and three anonymous BEA-9 reviewers for their helpful comments.

References

BNC. 2007. The British National Corpus, version 3. Distributed by Oxford University Computing Services on behalf of the BNC Consortium, `http://www.natcorp.ox.ac.uk/`.

Sergey Brin, James Davis, and Hector Garcia-Molina. 1995. Copy detection mechanisms for digital documents. In *Proceedings of the ACM SIGMOD Annual Conference*, pages 398–409.

Lei Chen and Su-Youn Yoon. 2011. Detecting structural events for assessing non-native speech. In *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications, NAACL-HLT*, pages 38–45, Portland, OR. Association for Computational Linguistics.

Chien-Ying Chen, Jen-Yuan Yeh, and Hao-Ren Ke. 2010. Plagiarism detection using ROUGE and WordNet. *Journal of Computing*, 2(3):34–44.

Pauline Cullen, Amanda French, and Vanessa Jakeman. 2014. *The Official Cambridge Guide to IELTS*. Cambridge University Press.

ETS. 2012. *The Official Guide to the TOEFL® Test, Fourth Edition*. McGraw-Hill.

Andrew Finch, Young-Sook Hwang, and Eiichiro Sumita. 2005. Using machine translation evaluation techniques to determine sentence-level semantic equivalence. In *Proceedings of the Third International Workshop on Paraphrasing*, pages 17–24.

Alexander Haputmann. 2006. Automatic spoken document retrieval. In Ketih Brown, editor, *Encylclopedia of Language and Linguistics (Second Edition)*, pages 95–103. Elsevier Science.

Timothy C. Hoad and Justin Zobel. 2003. Methods for identifying versioned and plagiarised documents. *Journal of the American Society for Information Science and Technology*, 54:203–215.

Claudia Leacock and Martin Chodorow. 1998. Combining local context and WordNet similarity for word sense identification. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 305–332. MIT Press.

Pearson Longman. 2010. *The Official Guide to Pearson Test of English Academic*. Pearson Education ESL.

Caroline Lyon, Ruth Barrett, and James Malcolm. 2006. Plagiarism is easy, but also easy to detect. *Plagiary*, 1:57–65.

Nitin Madnani, Joel Tetreault, and Martin Chodorow. 2012. Re-examining machine translation metrics for paraphrase identification. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 182–190, Montréal, Canada, June. Association for Computational Linguistics.

Thade Nahnsen, Özlem Uzuner, and Boris Katz. 2005. Lexical chains and sliding locality windows in content-based text similarity detection. CSAIL Technical Report, MIT-CSAIL-TR-2005-034.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Martin Potthast, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso. 2010. An evaluation framework for plagiarism detection. In *Proceedings of the 23rd International Conference on Computational Linguistics*.

Martin Potthast, Matthias Hagen, Tim Gollub, Martin Tippmann, Johannes Kiesel, Paolo Rosso, Efstathios Stamatatos, and Benno Stein. 2013. Overview of the 5th International Competition on Plagiarism Detection. In Pamela Forner, Roberto Navigli, and Dan Tufis, editors, *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers*.

Narayanan Shivakumar and Hector Garcia-Molina. 1995. SCAM: A copy detection mechanism for digital documents. In *Proceedings of the Second Annual Conference on the Theory and Practice of Digital Libraries*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Matt Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2008. TERp: A system description. In *Proceedings of the First NIST Metrics for Machine Translation Challenge (MetricsMATR)*, Waikiki, Hawaii, October.

Özlem Uzuner, Boris Katz, and Thade Nahnsen. 2005. Using syntactic information to identify plagiarism. In *Proceedings of the 2nd Workshop on Building Educational Applications using NLP. Ann Arbor*.

Zhibiao Wu and Martha Palmer. 1994. Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Mmeeting of the Association for Computational Linguistics (ACL)*.