Application of an Automatic Plagiarism Detection System in a Large-scale Assessment of English Speaking Proficiency

Xinhao Wang[1], **Keelan Evanini**[2], **Matthew Mulholland**[2], **Yao Qian**[1], **James V. Bruno**[2]
Educational Testing Service
[1]90 New Montgomery St #1450, San Francisco, CA 94105, USA
[2]660 Rosedale Road, Princeton, NJ 08541, USA
{xwang002, kevanini, mmulholland, yqian, jbruno}@ets.org

Abstract

This study aims to build an automatic system for the detection of plagiarized spoken responses in the context of an assessment of English speaking proficiency for non-native speakers. Classification models were trained to distinguish between plagiarized and non-plagiarized responses with two different types of features: text-to-text content similarity measures, which are commonly used in the task of plagiarism detection for written documents, and speaking proficiency measures, which were specifically designed for spontaneous speech and extracted using an automated speech scoring system. The experiments were first conducted on a large data set drawn from an operational English proficiency assessment across multiple years, and the best classifier on this heavily imbalanced data set resulted in an F1-score of 0.761 on the plagiarized class. This system was then validated on operational responses collected from a single administration of the assessment and achieved a recall of 0.897. The results indicate that the proposed system can potentially be used to improve the validity of both human and automated assessment of non-native spoken English.

1 Introduction

Plagiarism of spoken responses has become a vexing problem in the domain of spoken language assessment, in particular, the evaluation of non-native speaking proficiency, since there exists a vast amount of easily accessible online resources covering a wide variety of topics that test takers can use to prepare responses prior to the test. In the context of large-scale, standardized assessments of spoken English for academic purposes, such as the TOEFL iBT test (ETS, 2012), the Pearson Test of English Academic (Longman, 2010), and the IELTS Academic assessment (Cullen et al., 2014), some test takers may utilize content from online resources or other prepared sources in their spoken responses to test questions that are intended to elicit spontaneous speech. These responses that are based on canned material pose a problem for both human raters and automated scoring systems, and can reduce the validity of scores that are provided to the test takers; therefore, research into the automated detection of plagiarized spoken responses is necessary.

In this paper, we investigate a variety of features for automatically detecting plagiarized spoken responses in the context of a standardized assessment of English speaking proficiency. In addition to examining several commonly used text-to-text content similarity features, we also use features that compare various aspects of speaking proficiency across multiple responses provided by a test taker, based on the hypothesis that certain aspects of speaking proficiency, such as fluency, may be artificially inflated in a test taker's canned responses in comparison to non-canned responses. These features are designed to be independent of the availability of the reference source materials. Finally, we evaluate the effectiveness of this system on a data set with a large number of control (non-plagiarized) responses in an attempt to simulate the imbalanced distribution from an operational setting in which only a small number of the test takers' responses are plagiarized. In addition, we further validate this system on operational data and show how it can practically assist both human and automated scoring in a large scale assessment of English speaking proficiency

2 Previous Work

Previous research related to automated plagiarism detection for natural language has mainly focused on written documents. For example, a

Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 435–443
Florence, Italy, August 2, 2019. ©2019 Association for Computational Linguistics

series of shared tasks has enabled a variety of NLP approaches for detecting plagiarism to be compared on a standardized set of texts (Potthast et al., 2013), and several plagiarism detection services are available online.[1] Various techniques have been employed in this task, 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), stopword-based features (Stamatatos, 2011), features based on shared syntactic patterns (Uzuner et al., 2005), features based on word swaps detected via dependency parsing (Mozgovoy et al., 2007), and stylometric features (Stein et al., 2011), among others. In general, for the task of monolingual plagiarism detection, these methods can be categorized as either external plagiarism detection, in which a document is compared to a body of reference documents, or intrinsic plagiarism detection, in which a document is evaluated independently without a reference collection (Alzahrani et al., 2012). 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.

Little prior work has been conducted on the task of automatically detecting similar spoken responses, although research in the field of Spoken Document Retrieval (Hauptmann, 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.

In previous studies (Evanini and Wang, 2014; Wang et al., 2016), we conducted experiments on a simulated data set from an operational, large-scale, standardized English proficiency assessment and obtained initial results with an F1-measure of 70.6% using an automatic system to detect plagiarized spoken responses (Wang et al., 2016). Based on these previous findings, we extend this line of research and contribute in the following ways: 1) an improved automatic speech recognition (ASR) system based on Kaldi was introduced, and an unsupervised language model adaptation method was employed to improve the ASR performance on spontaneous speech elicited by new, unseen test questions; 2) an improved set of text-to-text content similarity features based on n-gram overlap and Word Mover's Distance was investigated; 3) in addition to evaluating the system on a simulated imbalanced data set, we also validated the developed automatic system using all of the responses from a single operational administration of the English speaking assessment in order to obtain a reliable estimate of the system's performance in a practical deployment.

3 Data

The data used in this study was drawn from a large-scale, high-stakes assessment of English for non-native speakers, which assesses English communication skills for academic purposes. The Speaking section of this assessment contains six tasks designed to elicit spontaneous spoken responses: two of them require test takers to provide an opinion or preference based on personal experience, which are referred to as independent tasks; and the other four tasks require test takers to summarize or discuss material provided in a reading and/or listening passage, which are referred to as integrated tasks.

In general, the independent tasks ask questions on topics that are familiar to test takers and are not based on any stimulus materials. Therefore, test takers can provide responses containing 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

[1] For example, http://turnitin.com/en_us/what-we-offer/feedback-studio,http://www.grammarly.com/plagiarism, and http://www.paperrater.com/plagiarism_checker.

and adapting it to the questions presented in the independent tasks. This type of plagiarism can affect the validity of a test taker's speaking score and can be grounds for score cancellation. 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 to test takers.

In order to identify the plagiarized spoken responses from the operational test, human raters need to first flag spoken responses that contained potentially plagiarized material, then trained experts subsequently review them to make the final decision. In the review process, the responses were transcribed and compared to external source materials 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 speaking task, the response was labeled as a plagiarized response and assigned a score of 0. In this study, a total of 1,557 plagiarized responses to independent test questions were collected from the operational assessment across multiple years.

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 responses in the data set along with a sample plagiarized response that apparently contains extended sequences of words directly matching idiosyncratic features of this source, such as the phrases "how romantic it can ever be" and "just relax yourself on the beach." In general, test takers typically do not reproduce the entire source material in their responses; rather, they often attempt to adapt the source material to a specific test question by providing some speech that is directly relevant to the prompt and combining it with the plagiarized material. An example of this is shown by the opening and closing non-italicized portions of the sample plagiarized response in Table 1. In total, human raters identified 224 different source materials while reviewing the potentially plagiarized responses, and their statistics information is as follows: the average number of words is 95.7 (std. dev. = 38.5), the average number of clauses is 10.3 (std. dev. = 5.1), and the average number of words per clause is 9.3 (std. dev. = 7.1).

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 multiple years, they were drawn from many different test forms, and it was not practical to obtain control data from all of the test forms that were represented in the plagiarized set. So, only the 166 test forms that appear most frequently in the canned data set were used for the collection of control responses, and 200 test takers were randomly selected from each form, without any overlap with speakers in the plagiarized set. The two spoken responses from the two independent questions were collected from each speaker; in total, 66,400 spoken responses from 33,200 speakers were obtained as the control set. Therefore, the data set used in this study is quite imbalanced: the number of control responses is larger than the number of plagiarized responses by a factor of 43.

4 Methodology

This study employed two different types of features in the automatic detection of plagiarized spoken responses: 1) similar to human raters' behavior in identifying the canned spoken responses, a set of features is developed to measure the content similarities between a test response and the source materials that were collected; 2) to deal with this particular task of plagiarism detection for spontaneous spoken responses, a set of features is introduced based on the assumption that the production of spoken language based on memorized material is expected to differ from the production of non-plagiarized speech in aspects of a test taker's speech delivery, such as fluency, pronunciation, and prosody.

4.1 Content Similarity

Previous work (Wang et al., 2016; Evanini and Wang, 2014) has demonstrated the effectiveness of using content-based features for the task of automatic plagiarized spoken response detection. Therefore, this study investigates the use of improved features based on the measurement of text-

Table 1: A sample source passage and the transcription of a sample plagiarized spoken response that was apparently drawn from the source. The test question/prompt used to elicit this response is also included. The overlapping word sequences between the source material and the transcription of the spoken response are indicated in italics.

Sample source passage: 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.	**Prompt:** Talk about an activity you enjoyed doing with your family when you were a kid. **Transcription of a plagiarized response:** family is a little trip to France when I was in primary school ten years ago I enjoy this activity first because we visited a small French town located by the beach the town *has very charming ocean view* and in *the sky 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* settling *down the sea birds are singing of course* I enjoy this activity with my family *also because there are many great French restaurants they offer the best sea food in the world like lobsters and tuna fishes* so I enjoy this activity with my family very much even it has passed several years

to-text content similarity. Given a test response, a comparison is made with each of the 224 reference sources using the following two content similarity metrics: word n-gram overlap and Word Mover's Distance. Then, the maximum similarity or the minimum distance is taken as a single feature to measure the content relevance between the test responses and the source materials.

4.1.1 N-gram Overlap

Features based on the BLEU metric have been proven to be effective in measuring the content appropriateness of spoken responses in the context of English proficiency assessment (Zechner and Wang, 2013) and in measuring content similarity in the detection of plagiarized spoken responses (Wang et al., 2016; Evanini and Wang, 2014). In this study, we first design a new type of feature, known as n-gram overlap, by simulating and improving the previous BLEU-based features. Word n-grams, with n varying from 1 word to 11 words, are first extracted from both the test response and each of the source materials, and then the number of overlapping n-grams are counted, where both n-gram types and tokens are counted separately. The intuition behind decreasing the maximum order is to increase the classifier's recall by evaluating the overlap of shorter word sequences, such as individual words in the unigram setting. On the other hand, the motivation behind increasing the maximum order is to boost the classifier's precision, since it will focus on matches of longer word sequences. Here, the maximum order of 11 was experimentally decided in consideration of the average number of words per clause in source materials, which is 9.3 as described in Section 3.

In order to calculate the maximum similarity across all source materials, the 11 n-gram overlap counts are combined together to generate one weighted score between a test and each source as in Equation 1, and then the maximum score across all sources is calculated as a feature to measure the similarity between a test and the set of source materials. Meanwhile, the 11 n-gram overlap counts calculated using the source with the maximum similarity score are also taken as features.

$$\sum_{i=1}^{n} \frac{i}{(n \cdot (n+1)/2)} \text{count\_overlap}(i\text{-gram}) \tag{1}$$

Furthermore, the n-gram based feature set can be enlarged by: 1) normalizing the n-gram counts by either the number of n-grams in the test response or the number of n-grams in each of the sources; 2) combining all source materials together into a single document for comparison (11 features based on n-gram overlap with the combined source), which is designed based on the assumption that test takers may attempt to use content from multiple sources. Similarly, this type of features can be further normalized by the number of n-grams in the test response. Based on all of these variations, a total of 116 n-gram overlap features is generated for each spoken response.

4.1.2 Word Mover's Distance

More recently, various approaches based on deep neural networks (DNN) and word-embeddings trained on large corpora have shown promising performance in document similarity detection (Kusner et al., 2015). In contrast to traditional similarity features, which are limited to a reliance on exact word matching, as the above n-gram overlap features, these new approaches have the advantage of capturing topically relevant words that are not identical. In this study, we employ Word Mover's Distance (WMD) (Kusner et al., 2015) to measure the distance between a test response and a source material based on word-embeddings.

Embeddings of words are first represented as vectors, and then the distance between each word appearing in a test response and each word in a source can be measured using the Euclidean distance in the embedding space. WMD represents the sum of the minimum values among the Euclidean word distances between words in the two compared documents. This minimization problem is a special case of Earth Mover's Distance (Rubner et al., 1998), for which efficient algorithms are available. Kunsner et al. (Kusner et al., 2015) reported that WMD outperformed other distance measures on document retrieval tasks and that the embeddings trained on the Google News corpus consistently performed well across a variety of contexts. For this work, we used the same word embeddings used in weighted embedding features as the input for the WMD calculation.

4.2 Difference in Speaking Proficiency

The performance of the above content similarity features greatly depends on the availability of a comprehensive set of source materials. If a test taker uses unseen source materials as the basis for a plagiarized response, the system may fail to detect it. Accordingly, a set of features that do not rely on a comparison with source materials has been proposed previously (Wang et al., 2016). The current study also examined this type of features.

As described in Section 3, the Speaking section of the assessment includes both independent and integrated tasks. In a given test administration, test takers are required to respond to all six questions; however, plagiarized responses are more likely to appear in the two independent tasks, since they are not based on specific reading and/or listening passages and thus elicit a wider range of variation across responses. Since the plagiarized responses are mostly constructed based on memorized material, they may be delivered in a more fluent and proficient manner compared to the responses that contain fully spontaneous speech. Based on this assumption, a set of features can be developed to capture the difference between various speaking proficiency features extracted from the canned and the fully spontaneous speech produced by the same test taker, where an automated spoken English assessment system, SpeechRater® (Zechner et al., 2007, 2009), can be used to provide the automatic proficiency scores along with 29 features measuring fluency, pronunciation, prosody, rhythm, vocabulary, and grammar. Since most plagiarized responses are expected to occur in the independent tasks, we assume the integrated responses are based on spontaneous speech. A mismatch between the proficiency scores and the feature values from the independent responses and the integrated responses from the same speaker can potentially indicate the presence of both prepared speech and spontaneous speech, and, therefore, the presence of plagiarized spoken responses.

Given an independent response from a test taker, along with the other independent response and four integrated responses from the same test taker, 6 features can be extracted according to each of the proficiency scores and 29 SpeechRater features. First, the difference of score/feature values between two independent responses was calculated as a feature, which was used to deal with the case in which only one independent response was canned and the other one contained spontaneous speech. Then, basic descriptive statistics, including mean, median, min, and max, were obtained across the four integrated responses. The differences between the score/feature value of the independent response and these four basic statistics were extracted as additional features. Finally, another feature was also extracted by standardizing the score/feature value of the independent response with the mean and standard deviation from the integrated responses. In total, a set of 180 features were extracted, referred as *SpeechRater* in the following experiments.

5 Experiments and Results

5.1 ASR Improvement

In this study, spoken responses need to be transcribed into text so that they can be compared with the source materials to measure the text-to-text similarity. However, due to the large amount of spoken responses considered in this study, it is not practical to manually transcribe all of them; therefore, a Kaldi[2]-based automatic speech recognition engine was employed. The training set used to develop the speech recognizer consists of similar responses (around 800 hours of speech) drawn from the same assessment and do not overlap with the data sets included in this study.

When using an ASR system to recognize spoken responses from new prompts that are not seen in the ASR training data, a degradation in recognition accuracy is expected because of the mismatch between the training and test data. In this study, we used an unsupervised language model (LM) adaptation method to improve the ASR performance on unseen data. In this method, two rounds of language model adaptation were conducted with the following steps: first, out-of-vocabulary (OOV) words from the prompt materials were added to the pronunciation dictionary and the baseline models were adapted with the prompts; second, the adapted models were applied to spoken responses from these new prompts to produce the recognized texts along with confidence scores corresponding to each response; third, automatically transcribed texts with confidence scores higher than a predefined threshold of 0.8 were selected; finally, these high-confident recognized texts were used to conduct another round of language model adaptation. We evaluated this unsupervised language model adaptation method on a stand-alone data set with 1,500 responses from 250 test speakers, where the prompts used to elicit these responses were unseen in the baseline recognition models. In this experiment, supervised language model adaptation with human transcriptions was also examined for comparison. As shown in Table 2, the overall word error rate (WER) is 25.5% for the baseline models. By applying the unsupervised LM adaptation method, the overall WER can be reduced to 22.9%; in comparison, the overall WER can be reduced to 21.8% by using supervised LM adaptation. The unsupervised method can achieve very similar results to the supervised method especially for responses to the independent prompts, i.e., with WER of 21.5% (unsupervised) vs 21.2% (supervised). These results indicate the effectiveness of the proposed unsupervised adaptation method, which was employed in the subsequent automatic plagiarism detection work.

Table 2: Word error rate (WER %) reduction with an unsupervised language model adaptation method, where the WERs on Independent items (IND), Integrated items (INT), as well as all items (ALL), are reported. The WER with the supervised adaptation method based on human transcriptions is also listed for comparison.

ASR Systems	IND	INT	ALL
Baseline	22.5	26.6	25.5
Unsupervised	21.5	23.5	22.9
Supervised	21.2	22.1	21.8

5.2 Experimental Setup

Due to ASR failures, a small number of responses were excluded from the experiments; finally, a total of 1,551 canned and 66,257 control responses were included in the simulated data. Since this work was conducted on a very imbalanced data set and only 2.3% of the responses in the simulated data are authentic plagiarized ones confirmed by human raters, 10-fold cross-validation was performed first on the simulated data. Afterward, the classification model built on the simulated data set was further evaluated on a corpus with real operational data.

We employed the machine learning tool of scikit-learn[3] (Pedregosa et al., 2011), for training the classifier. It provides various classification methods, such as decision tree, random forest, AdaBoost, etc. This study involves a variety of features from two different categories, and preliminary experiments demonstrated that the random forest model can achieve the overall better perfor-

[2] http://kaldi-asr.org/

[3] SKLL, a python tool making the running of scikit-learn experiments simpler, was used. Downloaded from `https://github.com/EducationalTestingService/skll`.

mance. Therefore, the random forest method is used to build classification models in the following experiments, and the precision, recall, as well as F1-score on the positive class (plagiarized responses) are used as the evaluation metrics.

5.3 Experimental Results

First, in order to verify the effectiveness of the newly developed n-gram overlap features, a preliminary experiment was conducted to compare this set of features with BLEU-based features, since they had been shown to be effective in previous research (Wang et al., 2016). The results as shown in Table 3 indicate that the F1-Measure of the n-gram features outperforms the BLEU features (0.761 vs. 0.748), and the recall of the n-gram features is higher than the BLEU features (0.716 vs. 0.683); inversely, the BLEU features result in higher precision (0.83 vs. 0.814). Accordingly, the n-gram features are used to replace the BLEU ones, since it is more important to reduce the number of false negatives, i.e., improve the recall, for our task.

Table 3: Comparison of n-gram and BLEU features.

Features	Precision	Recall	F1
BLEU	0.83	0.683	0.748
n-gram	0.814	0.716	0.761

Furthermore, each individual type of feature and their combinations were examined in the classification experiments described above. As shown in Table 4, each feature set alone was used to build classification models. The n-gram overlap features result in the best performance with an F1-score of 0.761, and the WMD features capturing the topical relevance between word pairs result in a much lower F1-score of 0.649. Furthermore, the combination of both types of content similarity features, i.e., n-gram and WMD, slightly reduces the F1-score to 0.76. These results indicate that for this particular task, the exact match of certain expressions appearing in both the test response and a source material plays a critical role.

As to the speaking proficiency related features, they can lead to a promising precision of 0.8 but with a very low recall of 0.009. After reexamining the assumption that human experts may be able to differentiate prepared speech from fully spontaneous speech based on the way how the speech is delivered, it turns out that it is quite challenging for human experts to make a reliable judgment of plagiarism just based on the speech delivery without any reference to the source materials, in particular, within the context of high-stakes language assessment. Accordingly, the features capturing the difference in speaking proficiency of prepared and spontaneous speech can be used as contributory information to improve the accuracy of an automatic detection system, but they are unable to achieve promising performance alone. Also as shown in Table 4, By adding the speaking proficiency features, the precision can be improved to 0.821, but the recall is reduced; finally, the F1-score is reduced to 0.752.

Table 4: Classification performance using each individual feature set and their combinations.

Features	Precision	Recall	F1
1. n-gram	0.814	0.716	0.761
2. WMD	0.663	0.636	0.649
3. SpeechRater	0.8	0.009	0.018
1 + 2	0.812	0.716	0.76
1 + 2 + 3	0.821	0.696	0.752

6 Employment in Operational Test

6.1 Operational Use

In order to obtain a more accurate estimate of how well the automatic plagiarism detection system might perform in a practical application in which the distribution is expected to be heavily skewed towards the non-plagiarized category, all test-taker responses to independent prompts were collected from a single administration of the speaking assessment. In total, 13,516 independent responses from 6,758 speakers were extracted for system evaluation. We collected 39 responses confirmed as plagiarized through the human review process, which represents 0.29% of the data set.

In this particular task, automatic detection systems can be applied to support human raters, where all potentially plagiarized responses can be first automatically identified and then human experts can be involved to review flagged responses and make the final decision about potential instances of plagiarism. In this scenario, it is more important to increase the number of true positives flagged by the automated system; thus, recall of plagiarized responses was used as the evaluation metric, i.e., how many responses can be successfully detected among these 39 confirmed instances of plagiarism.

6.2 Results and Discussion

In order to maximize the recall of plagiarized responses for this review, several models were built with either different types of features or different types of classification models, for example, random forest and Adaboost with decision tree as the weak classifier, and then they were combined in an ensemble manner to flag potentially plagiarized responses, i.e., a response is flagged if any of the models detects it as a plagiarized response. This ensemble system flagged 850 responses as instances of plagiarism in total and achieved a recall of 0.897, i.e., 35 of the confirmed plagiarized responses were successfully identified by the automatic system and 4 of them were missed.

These results prove that the developed system can provide a substantial benefit to both human and automated assessment of non-native spoken English. Manual identification of plagiarized responses can be a very difficult and time-consuming task, where human experts need to memorize hundreds of source materials and compare them to thousands of responses. By applying an automated system, potentially plagiarized responses can first be filtered out of the standard scoring pipeline; subsequently, experts can review these flagged responses to confirm whether they actually contain plagiarized material. Accordingly, instead of reviewing all 13,516 responses in this administration for plagiarized content, human effort is required only for the 850 flagged responses, thus substantially reducing the overall human effort. Thus, optimizing recall is appropriate in this targeted use case, since the number of false positives is within an acceptable range for the expert review. In addition, the source labels indicating which source materials were likely used in the preparation of each response are automatically generated by the automatic system for each suspected response; this information can help to accelerate the manual review process.

7 Conclusion and Future Work

This study proposed a system which can benefit a high-stakes assessment of English speaking proficiency by automatically detecting potentially plagiarized spoken responses, and investigated the empirical effectiveness of two different types of features. One is based on automatic plagiarism detection methods commonly applied to written texts, in which the content similarity between a test response and a set of source materials collected from human raters were measured. In addition, this study also adopted a set of features which do not rely on the human effort involved in source material collection and can be easily applied to unseen test questions. This type of feature attempts to capture the difference in speech patterns between prepared responses and fully spontaneous responses from the same speaker in a test. Finally, the classification models were evaluated on a large set of responses collected from an operational test, and the experimental results demonstrate that the automatic detection system can achieve an F1-measure of 0.761. Further evaluation on the real operational data also shows the effectiveness of the automatic detection system.

The task of applying an automatic system in a large-scale operational assessment is quite challenging since typically only a small number of plagiarized responses are distributed among a much larger amount of non-plagiarized responses to a wide range of different test questions. In the future, we will continue our research efforts to improve the automatic detection system along the following lines. First, since deep learning techniques have recently shown their effectiveness in the fields of both speech processing and natural language understanding, we will further explore various deep learning techniques to improve the metrics used to measure the content similarity between test responses and source materials. Next, further analysis will be conducted to determine the extent of differences between canned and spontaneous speech, and additional features will be explored based on the findings. In addition, another focus is to build automatic tools to regularly update the pool of source materials. Besides internet search, new sources can also be detected by comparing all candidate responses from the same test question, since different test takers may use the same source to produce their canned responses.

References

S. M. Alzahrani, N. Salim, and A. Abraham. 2012. Understanding plagiarism linguistic patterns, textual features, and detection methods. *IEEE Transaction on Systems, Man, and Cybernetics-Part C: Applications and Reviews*, 42(2):133–149.

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

C. Chen, J. Yeh, and H. Ke. 2010. Plagiarism detection using ROUGE and WordNet. *Journal of Computing*, 2(3):34–44.

P. Cullen, A. French, and V. Jakeman. 2014. *The Official Cambridge Guide to IELTS*. Cambridge University Press.

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

Keelan Evanini and Xinhao Wang. 2014. Automatic detection of plagiarized spoken responses. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–27.

A. Finch, Y. Hwang, and E. 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.

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

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

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 957–966, Lille, France.

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

C. Lyon, R. Barrett, and J. Malcolm. 2006. Plagiarism is easy, but also easy to detect. *Plagiary*, 1:57–65.

N. Madnani, J. Tetreault, and M. 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. Association for Computational Linguistics.

M. Mozgovoy, T. Kakkonen, and E. Sutinen. 2007. Using natural language parsers in plagiarism detection. In *Proceedings of the ISCA Workshop on Speech and Language Technology in Education (SLaTE)*.

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

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

M. Potthast, M. Hagen, T. Gollub, M. Tippmann, J. Kiesel, P. Rosso, E. Stamatatos, and B. Stein. 2013. Overview of the 5th International Competition on Plagiarism Detection. In *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers*.

M. Potthast, B. Stein, A. Barrón-Cedeño, and P. Rosso. 2010. An evaluation framework for plagiarism detection. In *Proceedings of the 23rd International Conference on Computational Linguistics*.

Y. Rubner, C. Tomasi, and L. J. Guibas. 1998. A metric for distributions with applications to image databases. In *Proceedings of the Sixth International Conference on Computer Vision*.

N. Shivakumar and H. 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*.

E. Stamatatos. 2011. Plagiarism detection using stopword n-grams. *American Society for Information Science and Technology*, 62(12):2512–2527.

B. Stein, N. Lipka, and P. Prettenhofer. 2011. Intrinsic plagiarism analysis. *Language Resources and Evaluation*, 45(1):63–82.

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

X. Wang, K. Evanini, J. Bruno, and M. Mulholland. 2016. Automatic plagiarism detection for spoken responses in an assessment of english language proficiency. In *Proceedings of the IEEE Spoken Language Technology Workshop*, pages 121–128.

K. Zechner, D. Higgins, and X. Xi. 2007. Speechrater^{SM}: A construct-driven approach to scoring spontaneous non-native speech. In *Proceedings of the International Speech Communication Association Special Interest Group on Speech and Language Technology in Education*, pages 128–131.

K. Zechner, D. Higgins, X. Xi, and D. M. Williamson. 2009. Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*, 51(10):883–895.

K. Zechner and X. Wang. 2013. Automated content scoring of spoken responses in an assessment for teachers of english. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, page 73–81.