Applying Unsupervised Learning To Support Vector Space Model Based Speaking Assessment

Lei Chen
Educational Testing Service
600 Rosedale Rd
Princeton, NJ
LChen@ets.org

Abstract

Vector Space Models (VSM) have been widely used in the language assessment field to provide measurements of students' vocabulary choices and content relevancy. However, training reference vectors (RV) in a VSM requires a time-consuming and costly human scoring process. To address this limitation, we applied unsupervised learning methods to reduce or even eliminate the human scoring step required for training RVs. Our experiments conducted on data from a non-native English speaking test suggest that the unsupervised topic clustering is better at selecting responses to train RVs than random selection. In addition, we conducted an experiment to totally eliminate the need of human scoring. Instead of using human rated scores to train RVs, we used used the machine-predicted scores from an automated speaking assessment system for training RVs. We obtained VSM-derived features that show promisingly high correlations to human-holistic scores, indicating that the costly human scoring process can be eliminated.

Index Terms: Vector Space Model (VSM), speech assessment, unsupervised learning, document clustering

1 Introduction

A Vector Space Model (VSM) is a simple, yet effective, method to measure similarities between documents or utterances, which has been utilized in the educational testing field. For example, VSM has been applied to detect students' off-topic essays (Higgins et al., 2006) and to automatically score essays (Attali and Burstein, 2004).

The following three steps are required to use VSM for automated assessment: (1) a collection of responses are selected from each score category to construct reference vectors (RV); (2) for an input response under scoring, the same vectorization method used for constructing RVs is applied to compute an input vector (IV); (3) similarities between this IV and the RVs for all score categories are computed as features reflecting vocabulary usage and content relevancy, including a widely used feature, the cosine similarity between the IV and the RV for the highest score category.

Clearly, the quality of VSM-derived features depends on the proper training of RVs. In language assessment, we tend to use a large number of manually scored responses to build RVs for each testing question (called *item* in the assessment field). However, this raises an issue: the requirement of manual scoring of these responses by human raters. Also, for large-scale assessments administrated globally, a high number of items are typically administered to both ensure the assessment security and support the large volume of test-takers. To address this challenge of application of VSM, we will describe our solutions based on applying unsupervised learning methods in this paper.

The rest of the paper is organized as follows: Section 2 reviews the related previous research; Section 3 describes the English assessment, the data used in our experiments, and the Automatic Speech Recognition (ASR) system used; Section 4 reports

Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pages 58–62,
Atlanta, Georgia, June 13 2013. ©2013 Association for Computational Linguistics

the three experiments we conducted; and Section 5 discusses our findings and plans for future research.

2 Previous Work

Attali and Burstein (2004) used the VSM method to measure non-native English writers' vocabulary choices when scoring their essays by comparing the words contained in an student's response to the words found in a sample of essays from each score category. One belief behind this methodology is that good essays will resemble each other in terms of the word choice. In particular, two VSM-derived features were used, including the maximum cosine similarity and cosine similarity to the top score category. Higgins et al. (2006) applied the VSM technology to detect students' off-topic essays whereby the word-based IV from a student's essay was compared to an RV built from a collection of on-topic essays. When the difference was larger than a pre-defined threshold, the essay was marked as off-topic. Zechner and Xi (2008) applied VSM as a content relevancy measurement to score non-native English speaking responses. Recently, Xie et al. (2012) explored the VSM technology on automated speech scoring. Using a superior ASR to the one used in (Zechner and Xi, 2008), they found that the VSM-derived features had moderately high correlations with human proficiency scores.

Dimension reduction, a critical step in applying VSM, removes the noises and minor details in word-based vectors and keeps a concise semantic structure. Latent Semantic Analysis (LSA) (Deerwester et al., 1990) and Latent Dirichlet Allocation (LDA) (Blei et al., 2003) are two widely used dimension-reduction methods. Kakkonen et al. (2005) systematically investigated the dimension reduction methods used in the VSM methods for essay grading. Their experiments showed that LSA slightly out-performs LDA.

Compared to supervised learning, unsupervised learning can skip the time-consuming and costly manual labeling process and has been widely used in many machine-learning tasks. Both LSA and LDA have been utilized in unsupervised document clustering (Hofmann, 2001) to automatically separate a collection of documents into several sets without any human intervention. Co-training is a type of semi-supervised learning method (Blum and Mitchell, 1998), consisting of two classifiers trained from independent sets of features to predict the same labels. It uses automatically predicted labels from one classifier to train the other classifier.

3 Data

The data used in our experiments were collected from the speaking section of Test Of English as a Foreign Language (TOEFL®), an English speaking test used to evaluate students' basic English-speaking skills for use in academic institutions that use English as their primary teaching language. Our data contains the speech responses for a total of 24 test items. For each item, both the stimulus material and question were presented to test-takers followed by a short amount of preparation time. The test-takers were then given up to 60 seconds to provide their spoken responses. These responses were scored by using carefully developed rating rubrics by a group of experienced human raters. The scoring rubrics covered a comprehensive list of different aspects of speaking ability, such as pronunciation, prosody, vocabulary, content organization, etc. A 4-point holistic scoring scale was used where the score of 4 marks the most advanced English speakers in the TOEFL® test. Table 1 summarizes the responses across these 24 items, including $mean$, sd, and sample size (n) of the total number of responses and the number of responses per each score level.

	Overall	$SC1$	$SC2$	$SC3$	$SC4$
$mean$	1969.63	81.88	701.96	963.46	222.33
sd	12.92	30.02	62.36	67.24	37.79
n	47271	1965	16847	23123	5336

Table 1: Summary statistics of the number of total responses and the number of responses per each score level measured in $mean$, sd, and sample size n across 24 items

The transcriptions of these spoken responses were obtained by running a state-of-the-art non-native ASR system. This ASR system uses a cross-word tri-phone acoustic model (AM) and n-gram language models (LMs) that were trained on approximately 800 hours of spoken data and the corresponding transcriptions. When being evaluated on an held-out data set transcribed by humans from the same test, a 33.0% word error rate was obtained.

4 Experiments

The three experiments described below shared the same procedure: (1) for each item, available responses were divided into two sets - a set for training RVs and a set for evaluating the VSM-derived features; (2) RVs were trained by using different response selection methods investigated in this paper; (3) the trained RVs were used to compute the VSM-derived features; and (4) Pearson correlation coefficients (rs) between the VSM-derived features and human-holistic scores were computed to measure these features' predictive abilities in speech scoring. This experimental procedure was conducted on all 24 items and was repeated in 10 iterations by using varied training/evaluation-splitting plans and the averages of these results across the items and iterations are reported. Note that we removed some common function words, such as *a*, *the*, etc., and some noise words from ASR outputs, such as *uh* and *um*, when applying the VSM method and always used LSA dimension reduction. We used the Gensim (Řehůřek and Sojka, 2010) Python package to implement the VSM-related computations in this paper. Also, in this paper, we focused on one VSM-derived feature $cos4$, the cosine distance between an IV to the RV representing the highest-score category (4) for TOEFL® test.

4.1 Data size for training RVs

In previous studies, researchers typically used a large number of responses to construct RVs. For example, Zechner and Xi (2008) used $1,000$ responses while Xie et al. (2012) increased the RV training data to $2,000$ responses for each item. We ask, is it possible to use fewer responses so that we would not be forced to manually score so many responses? To answer this question, we have investigated the relationship between the size of the RV training data and $cos4$'s predictive ability.

For each item, we first randomly selected $1,800$ responses as the RV training data and used the remaining responses as the evaluation set. We then gradually reduced the RV training set to $1,000$, 500, 200, and even 50 responses and trained a series of RVs. On the evaluation set, using these trained RVs, we extracted $cos4$ VSM feature and calculated the r_{cos4} for human-holistic scores. Table 2 reports the average r_{cos4}, which will de denoted as $\overline{r_{cos4}}$ thereafter, for the different-sized RV training sets. Table 2 shows that $\overline{r_{cos4}}$ continuously increases with the increase of the dataset size for training RVs. However, it is worth noting that using just 50 responses to train RVs still provides a reasonably high $\overline{r_{cos4}}$ (0.383). Between the two $size_{RV}$ conditions: 200 vs. 1800, $\overline{r_{cos4}}$ did not show a statistically significant increase based on a t-test ($p = 0.314$).

$size_{RV}$	50	200	500	1000	1800
$\overline{r_{cos4}}$	0.383	0.428	0.435	0.439	0.440

Table 2: $\overline{r_{cos4}}$, a measurement of VSM features' scoring performance, from different RV training data sizes

4.2 Using document clustering for training RVs

In the experiment described in section 4.1, we found that using even a limited number of human-scored responses can provide useful VSM features with a reasonably high r to human-holistic scores. If we can intelligently select such a small-sized dataset, we think that the VSM-derived features will show further improved predicting power. Armed with this idea, we proposed a solution to use unsupervised document clustering technology to find the responses for training RVs.

In particular, for each item, of the $1,800$ responses used for training the RVs, we run an LDA document-clustering process to split all of responses into K clusters. Then, for each cluster, we randomly selected M responses. Therefore, we selected $K \times M$ responses for human scoring and for training the RVs. Note that $K \times M$ can be much smaller than the original dataset size ($n = 1800$). We believed that comprehensive coverage of all of the latent topics would produce a better VSM that, in turn, would provide more effective VSM-derived features for scoring.

In our experiment, based upon a pilot study, we decided to use $K = 10$ and $M = 5$ to control the total scoring demand to be 50 responses per item. Compared to the $\overline{r_{cos4}}$ value obtained from randomly selecting 50 responses for training RVs (0.383 in Table 2), the response selection based on the document clustering improved the $\overline{r_{cos4}}$ to be 0.411. Furthermore, a t-test showed that such an increase in $\overline{r_{cos4}}$ is statistically significant ($p < 0.05$).

4.3 Using machine predicted scores for training RVs

Many of the previous automated speaking scoring systems focused on the features measuring fluency, pronunciation, and prosody (Witt, 1999; Franco et al., 2010; Bernstein et al., 2010; Chen et al., 2009). The scores predicted by these systems show promisingly high correlations with human rated scores. In order to eliminate the time-consuming and costly human scoring step required by applications of VSM, we considered using the scores automatically scored by algorithms (AS) instead of the scores rated by humans (HS).

In our experiment, we used a set of speech features following (Chen et al., 2009) for automated speech scoring. To estimate AS, a five-fold cross-validation was applied on the entire dataset. For each fold, a linear regression model was trained from 80% of responses by using their HS and was used to predict regression results on the remaining 20% of responses. The continuous scores produced by the regression model were rounded to the four discrete score levels (1 to 4) to serve as AS. Between the obtained AS and HS, a Pearson r 0.56 was observed.

Using the predicted scores, we re-ran our VSM feature experiment by using the $1,800$ responses to train the RVs. When the dataset sizes for training the RVs was at $1,800$, we found that the $\overline{r_{cos4}}$ was 0.410 when using machine-predicted scores. Although it was lower than the $\overline{r_{cos4}}$ value obtained by using human-rated scores (0.440), a feature with such correlational magnitude is still useful for building an automatic scoring model.

4.4 A summary of experiments

	HS_{1800}	HS_{50}	$HS_{cluster50}$	AS_{1800}
$\overline{r_{cos4}}$	0.440	0.383	0.411	0.410

Table 3: A summary of $\overline{r_{cos4}}$ using different RV training sizes, unsupervised-response clustering, and automated-predicted scores

Table 3 summarizes the three experiments described above. HS_{1800} refers to using $1,800$ responses with human scores (HS) to train RVs for each item. HS_{50} refers to using only 50 responses with human rated scores. $HS_{cluster50}$ refers to using 50 responses that were selected to cover 10 latent topics detected by using an LDA unsupervised topic clustering method. Compared to HS_{50}, we find that the unsupervised topic clustering method helped to improve $\overline{r_{cos4}}$. AS_{1800} refers to using $1,800$ responses with automatically predicted scores (AS) to train RVs for each item. Compared to HS_{1800}, AS_{1800} that avoids using a time-consuming and costly human scoring process, shows a reasonably high $\overline{r_{cos4}}$.

5 Conclusions and Future Work

Vector Space Models (VSMs) have been widely used in essay and speech assessment tasks to provide vocabulary usage and content relevance measurements. However, applying VSM on the assessments with many items requires a lot of work by human raters. To make the application of VSM in assessments more economical and efficient, we propose the use of unsupervised learning methods to reduce and even eliminate the time-consuming and costly human-scoring process. First, we found that it was possible to just use hundreds rather than thousands of responses to train RVs when applying VSM. In our experiments with TOEFL® data, we found that using a minimum 200 responses to train RVs for each item, was not statistically significantly different from using $1,800$ responses. Next, we used an LDA document-clustering method to identify latent topics from all of the items and used the topic information to select responses for training RVs. Our experiments clearly suggest that such a method of selection provides more effective VSM features than random selection. Finally, we used the scores predicted by an automated speech scoring system that mostly uses fluency and pronunciation features to replace human-rated scores in building the VSM. Our experiments suggest that the features derived from such a VSM that can be constructed without the need of human scoring show promisingly high correlations to human-holistic scores.

This research can be extended in several new directions. First, we will apply the proposed methods on other language assessment tasks, such as on long (written) essays, to fully test that the proposed methods are universally helpful. Second, we are considering doing the third experiment in more iterations – adding the VSM-derived features into the automated scoring model so that more accurate machine-predicted scores can be used for building further improved VSM.

References

Y. Attali and J. Burstein. 2004. Automated essay scoring with e-rater v.2.0. In *Presented at the Annual Meeting of the International Association for Educational Assessment*.

J. Bernstein, A. Van Moere, and J. Cheng. 2010. Validating automated speaking tests. *Language Testing*, 27(3):355.

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

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, page 92100.

L. Chen, K. Zechner, and X Xi. 2009. Improved pronunciation features for construct-driven assessment of non-native spontaneous speech. In *NAACL-HLT*.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391407.

H. Franco, H. Bratt, R. Rossier, V. Rao Gadde, E. Shriberg, V. Abrash, and K. Precoda. 2010. EduSpeak: a speech recognition and pronunciation scoring toolkit for computer-aided language learning applications. *Language Testing*, 27(3):401.

D. Higgins, J. Burstein, and Y. Attali. 2006. Identifying off-topic student essays without topic-specific training data. *Natural Language Engineering*, 12.

Thomas Hofmann. 2001. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1):177–196.

Tuomo Kakkonen, Niko Myller, Erkki Sutinen, and Jari Timonen. 2005. Comparison of dimension reduction methods for automated essay grading. *Natural Language Engineering*, 1:1–16.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA.

S. M. Witt. 1999. *Use of Speech Recognition in Computer-assisted Language Learning*. Ph.D. thesis, University of Cambridge.

S. Xie, K. Evanini, and K. Zechner. 2012. Exploring content features for automated speech scoring. *Proceedings of the NAACL-HLT, Montreal, July*.

Klaus Zechner and Xiaoming Xi. 2008. Towards automatic scoring of a test of spoken language with heterogeneous task types. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 98–106. Association for Computational Linguistics.