Using Ontology-based Approaches to Representing Speech Transcripts for Automated Speech Scoring

Miao Chen
School of Information Studies
Syracuse University
Syracuse, NY 13244, USA
mchen14@Syr.edu

Abstract

This paper presents a thesis proposal on approaches to automatically scoring non-native speech from second language tests. Current speech scoring systems assess speech by primarily using acoustic features such as fluency and pronunciation; however content features are barely involved. Motivated by this limitation, the study aims to investigate the use of content features in speech scoring systems. For content features, a central question is how speech content can be represented in appropriate means to facilitate automated speech scoring. The study proposes using ontology-based representation to perform concept level representation on speech transcripts, and furthermore the content features computed from ontology-based representation may facilitate speech scoring. One baseline and two ontology-based representations are compared in experiments. Preliminary results show that ontology-based representation slightly improves performance of one content feature for automated scoring over the baseline system.

1 Introduction

With increasing number of language learners taking second language tests, the resulting responses add a huge burden to testing agencies, and thus automated scoring has become a necessity for efficiency and objectivity. Speaking, an important aspect for assessing second language speakers' proficiency, is selected as the context of the study. The general goal is to investigate new approaches to automatic scoring of second language speech.

When giving a speaking test in computer-mediated environment, test-takers' responses are typically recorded as speech files. These files can be considered to contain two layers: sound and text. The sound is about the acoustic side of speech, whose features have been used to assess speaking proficiency in existing automated speech-scoring systems (Dodigovic, 2009; Zechner et al., 2009). However, the text side, which is about the content of speech, is by far not well addressed in scoring systems, mainly due to the imperfect performance of automatic speech recognizer systems. As content is an integral part of speech, adding content features to existing scoring systems may further enhance system performance, and thus this study aims to examine the use of content features in speech scoring systems.

In order to acquire speech content, speech files need to be transcribed to text files, by human or Automatic Speech Recognition (ASR). The resulted text files, namely, speech transcripts, are to be processed to extract content features. Moreover, representation of text content (e.g. in vectors) is important because it is the pre-requisite for computing content features and building speech scoring models. Therefore this study focuses on representing content of speech transcripts to facilitate automatic scoring of speech.

Speech transcripts can be seen as a special type of text documents, and therefore document representation approaches shed light on representation of speech transcripts, such as Salton et al. (1975),

Proceedings of the NAACL HLT 2012 Student Research Workshop, pages 41–47,
Montréal, Canada, June 3-8, 2012.

Deerwester et al. (1990), Lewis (1992), Kaski (1997), He et al. (2004), Arguello et al. (2008), Hotho et al. (2003a). On the other hand, written essays, the output of writing section of second language test, share great similarity with speech transcripts, and representation of essays also has implications on speech transcript representation, such as Burstein (2003), Attali & Burstein (2006), and Larkey & Croft (2003).

Existing document representation approaches are primarily statistical and corpus based, using words or latent variables mined from corpus as representation units in the vector. These approaches exhibit two challenges: 1) meaningfulness of representation units. For example, synonymous words represent similar meaning and thus should be grouped as one representation unit. 2) unknown terms. Since words or latent variables in the vector are from training corpus, if an unknown term occurs in the testing corpus then it is difficult to determine the importance of the term in the training corpus because there is no prior knowledge of it in the training corpus.

Ontology concepts, representation units at the concept level, have been less employed in content representation. Hotho et al. (2003a) claim that ontology concepts can help reveal concepts and semantics in documents, and thus we hypothesize ontology-based representation may facilitate obtaining better content features for speech scoring. Ontologies can also complement the abovementioned shortcomings of statistical and corpus based representations by providing meaningful representation units and reasoning power between concepts.

The study compares baseline (statistical and corpus based) and ontology-based approaches. The criterion is representing the same speech transcripts using these approaches, computing content features based on the representations, and comparing performance of content features in predicting speaking proficiency.

2 Related Work

This section reviews work related to representation of speech transcripts, including document representation, essay scoring, and ontology-based representation in text processing.

Document representation has been an important topic in research areas such as natural language processing, information retrieval, and text mining, in which a number of representation approaches have been proposed.

The most common practice of text document representation is the Bag-Of-Words (BOW) approach, illustrated in Salton et al. (1975). The basic idea is that a document can be represented as a vector of words, with each dimension of the vector standing for one single word. Besides using explicit words from documents, latent variables derived from document mining can be used for document representation as well, such as the Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) approaches. The representation units are latent concepts or topics and the documents are projected to the semantic space constructed from the latent concepts or topics (Deerwester et al., 1990; Blei, 2012). An important purpose of using the latent variables is to reduce dimensions in document representation and place documents in a more compact space. Some other dimension reduction techniques include Subspace-based methods (Chen et al., 2003) and Self Organizing Map (Kaski 1997; Kaski, et al. 1998).

In the area of automatic essay scoring, essay content are represented to facilitate the scoring. The BOW approach is widely used in essay representation as well, including the e-rater system (Burstein, 2003; Attali & Burstein, 2006) and the experimental system in Larkey & Croft (2003). Representation in the BETSY system (Bayesian Essay Test Scoring System) also involves words, such as frequency of content words, along with specific phrases (Dikli, 2006). The exemplar system employing LSA representation is the Intelligent Essay Assessor system, which performs LSA on training essays and then projects essays to the vector space of latent concepts (Landauer et al., 2003).

Besides representation approaches, content feature computing in essay scoring is useful to content scoring of speech because they share great similarity. Content features can be derived by computing cosine similarity between essay vectors, such as in e-rater (Attali & Burstein, 2006) and Intelligent Essay Assessor (Landauer et al., 2003). The e-rater content feature computing is adopted in this study to compute content features of speech transcripts.

As mentioned in section 1, ontologies can be used to complement the challenges of statistical

representation approaches. Ontology concepts have been successfully used in text processing tasks such as text classification and clustering and can help resolve the first challenge, meaningfulness of the representation. Hotho and Bloehdorn along with other people conducted a series of studies in using the ontologies (i.e. WordNet) for text categorization and clustering tasks (Bloehdorn & Hotho, 2004; Hotho et al., 2003a; Hotho et al., 2003b). The goals are to overcome several weaknesses, e.g. synonyms and generalization issues, of the bag-of-words representation by using ontology concept based representation. Basically concepts from ontologies are used as units for text representation and then text processing is performed on top of the ontology-based representation. Explicit Semantic Approach (ESA), proposed by Gabrilovich and Markovitch (2007), is an approach to representing an arbitrary text snippet in vector of Wikipedia concepts for the convenience of further natural language processing. Each Wikipedia concept has text description, which is used to build an invert index to associate words with concepts. The invert index helps represent each word by vector of Wikipedia concepts, and eventually a document can be represented by weighted Wikipedia concepts by adding up the Wikipedia concept vectors of the words that the document contains.

Ontologies can also help resolve the second challenge of statistical representation, the unknown term issue. If a term occurs in the testing corpus but not the training corpus, then the importance of the term can be inferred from external knowledge such as ontologies. The semantic relations defined in ontologies connect relevant concepts and organize them into a tree (i.e. WordNet) or a graph structure (i.e. Wikipedia). Since paths usually exist between two individual concepts, ontologies can support inferences among concepts by using the paths and concept nodes between them. Moreover, semantic similarity between concepts, computed based on ontology knowledge, can be used to infer importance of unknown terms.

WordNet (Fellbaum, 1998) and Wikipedia (Wikipedia, 2012) ontologies are two popular ontologies for computing semantic similarity. A number of similarity approaches have been proposed for similarity calculation according to the different characteristics of the two ontologies (Lin, 1998; Pedersen et al., 2004; Resnik, 1999; Strube & Ponzetto, 2006).

3 Methodology

Experiments are conducted to compare ontology-based representations (experimental systems) and common representations (baseline systems). Two ontology-based methods are employed as the experimental systems, one is about representing transcripts using ontology concepts ("ONTO"), and the other is about inferring weights of unknown terms using ontologies ("OntoReason"). For the baseline, we identify the BOW representation as a common text representation and use it in the baseline system.

3.1 Data Set

The data set comes from an English proficiency test for non-native speakers. For the speaking section, test takers are asked to provide spontaneous speech responses to the prompts[1] (test tasks). There are 4 prompts in the data set, all of which are integrated prompts. An integrated prompt is a test task that first provides test takers some materials to read or listen and then asks them to provide opinions or arguments towards the materials. The responses are then scored holistically by human raters based on a scoring rubric on a scale of 1 to 4, 4 being the highest score. For each score level, the scoring rubric contains guidelines of expected performance on various aspects of speaking ability such as pronunciation, fluency, and content.

The data set contains 1243 speech samples from 327 speakers in total. Manual and automatic methods are used to obtain transcripts of the speech samples. For the manual way, each response is verbatim transcribed by human; and for the automatic way, each response is automatically transcribed by ASR with word error rate of 12.8%. Therefore two sets of transcripts are derived for the speech responses, the human transcripts set and the ASR set.

Since the representation approaches are prompt-specific in the study, meaning vector representations are generated for each prompt, the data set is first split by prompts and then responses are split into training and testing sets within each prompt. Table 1 shows size of the data set and subsets:

[1] Prompts are test tasks assigned to test takers to elicit their speaking responses.

[2] The feature is referred to as "cos.w/6" in Attali and Burstein (2006) because there are usually 6 score levels, while here our

Prompt	Training Set	Test Set	Total
A	143	176	319 (4/79/158/78)
B	140	168	308 (7/86/146/69)
C	139	172	311 (4/74/154/79)
D	137	168	305 (8/75/141/81)

Table 1. Size of data set and subsets. The numbers in parentheses are the number of documents on score levels 1-4.

3.2 Representation Approaches of Speech Transcripts

One baseline approach and two ontology-based approaches are briefly introduced here and implemented in experiments. The approaches are used to generate vectors for computing content features. We also plan to employ other approaches in the future, as described in section 5.

3.2.1 Bag-of-words (baseline)

It takes the view that essays can be represented in vector of words and the value of a word in a vector refers to its weighting on this dimension. It uses the representation method in the e-rater as well, including document-level representation for testing documents and score-level representation for training documents (Attali & Burstein, 2006).

Within each prompt, each testing transcript is converted to a vector (document level representation); training transcripts are grouped by their score levels and for each score level a vector is generated by aggregating all transcripts of this score level (score level representation). We decide to use the tfidf weighting schema with stop words removed after tuning options of the parameters.

3.2.2 Ontology-based Representation (experimental)

ONTO-WordNet approach. Concepts from ontology are identified in speech transcripts and then used to generate concept-level vectors. In practice, concept mapping in transcripts varies according to characteristics of ontologies. The WordNet ontology, containing mostly single words, is used as one case in the study. In the future, we plan to try the Wikipedia ontology, which contains more phrases-based concepts, for ontology-based representation.

Synsets, groups of synonyms, are concepts in WordNet and used as ontology concepts here. Document text is split by whitespace and punctuations to a set of words. Then the words are matched to WordNet synsets. As a word may have multiple senses (synsets), it is necessary to decide which synset to use in WordNet. Therefore we try two sense selection strategies as in Hotho et al.'s (2003a) study: 1) simply use the first sense in WordNet; and 2) do part-of-speech tagging on sentences and find the corresponding sense in WordNet. We find the 1st strategy obtains better performance than the 2nd one and thus decide to use the 1st one. When constructing ontology-based vector, we include both concepts and words in the vector.

3.2.3 Ontology-based Reasoning (experimental)

OntoReason-WordNet approach. This approach is also implemented by using WordNet. First, transcripts are represented by ontology concepts as in section 3.2.2. Then given an unknown concept in test transcripts, we identify its semantically similar concepts (N=5) in the training transcripts and then reason the weight of the unknown concept based on the weights of these similar concepts.

The reasoning makes use of semantic similarity between WordNet synsets. Concept similarity is computed using the edge-based path similarity (Pedersen et al., 2004). We select N=5 concepts from the training transcripts that are most similar to the unknown concept, and compute the weight of the unknown concept in the training transcripts by averaging the weights of the 5 similar concepts.

3.3 Content Feature Computation

The baseline and experimental systems all generate vector representations for speech transcripts. The content features are computed based on vector representation, and all representation approaches employ the same method of computing content features. We choose to use the two content features of the e-rater system, "max.cos" and "cos.w4", as the feature computation method[2] (Attali & Burstein, 2006).

The max.cos feature. This feature identifies which score level of training documents the testing document is closest to. It computes and compares the similarity between the test document and training documents of each score level in vector space, and then makes the score level whose training doc-

[2] The feature is referred to as "cos.w/6" in Attali and Burstein (2006) because there are usually 6 score levels, while here our data has 4 score levels therefore it is written as "cos.w4".

uments are most similar to the test document as the feature value.

The cos.w4 feature. This feature computes content similarity between the test document and the highest level training documents in vector space. Since score 4 is the highest level in our data set of spoken responses, we compute the cosine similarity between the test vector and the score level 4 vector as the feature value.

Given a speech transcript from the test set, we first convert it to a vector using one of the representation approaches, and then compute the max.cos and cos.w4 feature values as its content features.

3.4 Evaluation

Representation approaches are evaluated based on their performance in predicting speaking proficiency of test takers. More specifically, a representation approach generates a vector representation using specific representation units (e.g. words, concepts); for each test transcript, two content features are computed based on the vector representation; Pearson correlation r is computed between each content feature and speaking proficiency to indicate the predictiveness of the content feature resulting from a specific representation. Higher correlation indicates higher predictiveness on speaking proficiency. Lastly, we compare content feature correlations of different representation approaches. We consider that the higher the correlation is, the better the representation approach is.

4 Experiment Results

In the preliminary stage, the BOW (baseline), ONTO-WordNet and OntoReason-WordNet (experimental) approaches are implemented. Meanwhile parameters are optimized to acquire the best parameter setup for each approach. Since the speech files are transcribed by both human and ASR, same experiments are run on both data sets to compare representation performance on different transcriptions. The correlations of the two content features to speaking proficiency are computed for each representation. Tables 2 and 3 show correlations of the max.cos and cos.w4 features respectively:

For the max.cos feature, the average correlation of the ONTO-WordNet approach outperforms the BOW baseline slightly but the correlation drops dramatically when using the OntoReason-WordNet approach, for both the human and ASR transcripts. For the cos.w4 feature, the average correlation of the ONTO-WordNet approach outperforms the BOW, and the OntoReason-WordNet further outperforms the ONTO-WordNet approach, for both the human and ASR transcripts. It shows some evidence that ontology-based representation can improve performance of both content features; the ontology-based reasoning increases performance of the cos.w4 feature but decreases the max.cos feature correlation.

Comparing the performance on human vs. ASR transcripts, the features extracted from the human transcripts exhibit better average correlations than the corresponding features from the ASR transcripts. The results also show that the correlation difference between human and ASR transcripts is moderate. It may indicate that the representation approaches can be employed on ASR transcripts to further automate the speech scoring process.

Prompt	Hum, BOW	Hum, ONTO-WordNet	Hum, Onto-Reason-WordNet	ASR, BOW	ASR, ONTO-WordNet	ASR, Onto-Reason-WordNet
A	0.320	0.333	0.038	0.293	0.286	0.014
B	0.348	0.352	0.350	0.308	0.338	0.339
C	0.366	0.373	0.074	0.396	0.386	0.106
D	0.343	0.323	0.265	0.309	0.309	0.265
Average	0.344	0.345	0.182	0.327	0.330	0.181

Table 2. Correlations between the max.cos feature and speaking proficiency (Hum=using human transcriptions; ASR=using ASR hypotheses).

Prompt	Hum, BOW	Hum, ONTO-WordNet	Hum, Onto-Reason-WordNet	ASR, BOW	ASR, ONTO-WordNet	ASR, Onto-Reason-WordNet
A	0.427	0.429	0.434	0.409	0.416	0.411
B	0.295	0.303	0.327	0.259	0.278	0.292
C	0.352	0.385	0.402	0.338	0.366	0.380
D	0.368	0.385	0.389	0.360	0.379	0.374
Average	0.361	0.376	0.388	0.342	0.360	0.364

Table 3. Correlations between the cos.w4 feature and speaking proficiency (Hum=using human transcriptions; ASR=using ASR hypotheses)

5 Future Work

For future work, we will implement one more baseline (LSA) and two more ontology-based approaches (ONTO-Wikipedia and OntoReason-Wikipedia) and analyze their performance.

Latent semantic analysis (LSA). LSA decomposes a term-by-document matrix generated from training transcripts to three sub-matrices. Then given a test transcript, documents can be projected to the latent semantic space based on the three sub-matrices. The rank k parameter needs to be decided as a parameter for dimensionality reduction purpose by tuning it on the training data.

Using Wikipedia as another case for ontology, two more experimental approaches will be implemented, one for ontology-based representation and the other for ontology-based reasoning.

ONTO-Wikipedia. Wikipedia concepts can be identified in transcripts in two ways: 1) directly find concepts in text window of 5 words; 2) convert a transcript in vectors of Wikipedia concepts using the Explicit Semantic Analysis method, which associates words to Wikipedia concepts and represents arbitrary text using the word-concept associations (Gabrilovich and Markovitch, 2007).

OntoReason-Wikpedia. The concept similarity between Wikipedia concepts is obtained by computing the cosine similarity of the text description of the concepts. The reasoning method of the unknown concept follows the one mentioned in the OntoReason-WordNet approach.

We will compute content features based on these new representations and evaluate the performance according to feature correlations. The current results examine effects of using the WordNet ontology on predicting speaking proficiency, and these new experiments will answer whether the other type of ontology, Wikipedia, has positive effect in speaking proficiency prediction. We will also compare the effects of using different ontologies for ontology-based representations.

The study has implications on effects of different speech transcript representations in predicting speaking proficiency. Since content features are less well explored in automatic speech scoring compared to acoustic features, it also contributes to the understanding of the use and effects of content features in speech scoring.

Acknowledgments

The author would like to than Drs. Klaus Zechner and Jian Qin for their tremendous help and support on the dissertation study.

References

Arguello, J., Elsas, J. L., Callan, J., & Carbonell, J. G. (2008). *Document representation and query expansion models for blog recommendation*. Proceedings of the Second International Conference on Weblogs and Social Media (ICWSM 2008).

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. *The Journal of Technology, Learning and Assessment*, 4(3).

Blei, D. (2012). Introduction to probabilistic topic models. *Communications of the ACM, 77-84*.

Bloehdorn, S., & Hotho, A. (2004). *Boosting for text classification with semantic features.* Workshop on mining for and from the semantic web at the 10th ACM SIGKDD conference on knowledge discovery and data mining (KDD 2004).

Burstein, J. (2003). The E-rater® scoring engine: Automated essay scoring with natural language processing. In M. D. Shermis, Burstein, J.C. (Ed.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 113-121). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Chen, L., Tokuda, N., & Nagai, A. (2003). A new differential LSI space-based probabilistic document classifier. *Information Processing Letters*, 88(5), 203-212.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent

semantic analysis. *Journal of the American Society for information science*, 41(6), 391-407.

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, 5(1).

Dodigovic, M. (2009). *Speech Processing Technology in Second Language Testing*. Proceedings of the Conference on Language & Technology 2009.

Fellbaum, C. (Ed.). (1998). *WordNet: An electronic lexical database*. Cambridge, MA: The MIT press.

Gabrilovich, E., & Markovitch, S. (2007). *Computing semantic relatedness using wikipedia-based explicit semantic analysis*. Proceedings of the 20th International Joint Conference on Artificial Intelligence.

He, X., Cai, D., Liu, H., & Ma, W. Y. (2004). *Locality preserving indexing for document representation*. Proceedings of the 27th Annual International ACM SIGIR Conference.

Hotho, A., Staab, S., & Stumme, G. (2003a). *Ontologies improve text document clustering*. Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03).

Hotho, A., Staab, S., & Stumme, G. (2003b). *Text clustering based on background knowledge* (Technical report, no.425.): Institute of Applied Informatics and Formal Description Methods AIFB, University of Karlsruche.

Kaski, S. (1997). Computationally efficient approximation of a probabilistic model for document representation in the WEBSOM full-text analysis method. *Neural processing letters*, 5(2), 69-81.

Kaski, S., Honkela, T., Lagus, K., & Kohonen, T. (1998). WEBSOM-Self-organizing maps of document collections1. *Neurocomputing*, 21(1-3), 101-117.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis, Burstein, J.C. (Ed.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 87–112). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Larkey, L. S., & Croft, W. B. (2003). A Text Categorization Approach to Automated Essay Grading. In M. D. Shermis & J. C. Burstein (Eds.), *Automated Essay Scoring: A Cross-discipline Perspective*: Mahwah, NJ, Lawrence Erlbaum.

Lewis, D. D. (1992). *Representation and learning in information retrieval*. (Doctoral dissertation). University of Massachusetts.

Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004). *WordNet:: Similarity: measuring the relatedness of concepts*. Proceedings of the Fifth Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-04).

Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence*, 11(1999), 95-130.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11), 613-620.

Strube, M., & Ponzetto, S. P. (2006). *WikiRelated! Computing semantic relatedness using Wikipedia*. Proceedings of the American Association for Artificial Intelligence 2006, Boston, MA.

Wikipedia: The free encyclopedia. (2012, Apr 1). FL: Wikimedia Foundation, Inc. Retrieved Apr 1, 2012, from http://www.wikipedia.org

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