Using an Ontology for Improved Automated Content Scoring of Spontaneous Non-Native Speech

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

Klaus Zechner
Educational Testing Service
660 Rosedale Road
Princeton, NJ 08541, USA
kzechner@ets.org

Abstract

This paper presents an exploration into automated content scoring of non-native spontaneous speech using ontology-based information to enhance a vector space approach. We use content vector analysis as a baseline and evaluate the correlations between human rater proficiency scores and two cosine-similarity-based features, previously used in the context of automated essay scoring. We use two ontology-facilitated approaches to improve feature correlations by exploiting the semantic knowledge encoded in WordNet: (1) extending word vectors with semantic concepts from the WordNet ontology (synsets); and (2) using a reasoning approach for estimating the concept weights of concepts not present in the set of training responses by exploiting the hierarchical structure of WordNet. Furthermore, we compare features computed from human transcriptions of spoken responses with features based on output from an automatic speech recognizer. We find that (1) for one of the two features, both ontologically based approaches improve average feature correlations with human scores, and that (2) the correlations for both features decrease only marginally when moving from human speech transcriptions to speech recognizer output.

1 Introduction

Currently, automated speech scoring systems mainly utilize features related to the acoustic aspects of a spoken response of a test taker, for example, fluency, pronunciation, and prosody features (Cucchiarini et al., 2000, 2002; Franco et al., 2010; Zechner et al., 2009). In terms of the content aspect of speech, for highly predictable speech, such as reading a passage aloud, scoring of content reduces to measuring the reading accuracy of the read passage which is typically achieved by computing the string edit distance between the target passage and the actual text read by the test taker, using the speech recognizer hypothesis as a proxy (Alwan et al., 2007; Balogh et al., 2007). For high entropy speech whose content is difficult to predict such as spontaneous speech in this study, on the other hand, content scoring has not been investigated much so far, mostly due to the difficulty of obtaining accurate word hypotheses for spontaneous non-native speech by Automated Speech Recognition (ASR) systems.

In this paper, we use spoken responses from an English language spoken proficiency test where candidates, all non-native speakers of English, respond to four different prompts[1] with a speaking time of one minute per response.

For this study, we decide to use a baseline approach for content scoring of spontaneous speech that was previously employed for a similar task in the context of automated essay scoring (Attali & Burstein, 2006), namely Content Vector Analysis (CVA) where every document is represented as a vector of word weights, based on their frequencies in a document or document collection. However, there are two issues with the CVA vector of words representation that we want to address with this study: (1) Similar words are treated in isolation and not grouped together. Words with similar meaning should be treated in the same way in an automated scoring system, so grouping word synonyms into semantic concepts can help with this issue. (2) The vector of word representation is based on an exist-

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

The 7th Workshop on the Innovative Use of NLP for Building Educational Applications, pages 86–94,
Montréal, Canada, June 3-8, 2012. ©2012 Association for Computational Linguistics

ing corpus of training documents. When encountering a word or concept in a test document that is not contained in the training set, it is difficult to decide the relevance of that word or concept.

We propose to use ontology-facilitated approaches as solutions to these two issues, aiming at enriching speech content representations to improve speech content scoring. Specifically, to address issue (1), we represent speech content by concept-level vectors, using the *synsets* (lists of synonymous words) of the WordNet ontology (Fellbaum, 1998; WordNet 3.0, 2010). As for issue (2), we expand the vector representation by inferring the importance (weight) of concepts not present in the training vectors based on their path distance to known concepts or words in the hierarchical structure of the WordNet ontology.

Since we only look at the content aspect of speech without considering the acoustic features in this study, we work on speech transcripts exclusively, both from human transcribers as well as from a state-of-the-art automated speech recognition system, and compare results between the ideal human transcripts and the imperfect transcripts generated by the speech recognizer. For the purpose of simplified illustration, speech transcripts are often referred to as "documents" in the paper as they are a special type of textual documents.

The remainder of this paper is organized as follows: in Section 2, we review related research in content scoring of texts, particularly student essays; Section 3 describes the data set we use for this study and the ASR system; and Section 4 presents the ontologically-facilitated methods we are using in detail. In Section 5, we present our experiments along with their results, followed by a discussion in Section 6, and we conclude the paper with a summary and outlook in Section 7.

2 Related Work

There have been some effective approaches for test takers' written responses in language tests, namely in the area of Automated Essay Scoring (AES).

AES has employed content vector analysis, i.e., vectors of words to represent text, for example, the e-rater system (Burstein, 2003; Attali & Burstein, 2006) and the experimental system in Larkey and Croft (2003). Representations in the BETSY system (Bayesian Essay Test Scoring System) also involve words, such as the frequency of content words, and also include specific phrases as well (Dikli, 2006). AES has also used latent concepts for text representation, such as the Intelligent Essay Assessor system (Landauer et al., 2003). The latent concepts are generated by a statistical approach called Latent Semantic Analysis (LSA), which constructs a semantic vector space and projects essays to the new space.

Representing texts by vectors of words has also been a common practice in many research areas beyond AES, including information retrieval (Salton et al., 1975; Croft et al., 2010). One of its weaknesses, however, is its difficulty in addressing issues such as synonyms and related terms. Different words, such as lawyer, attorney, counsel etc. can share similar meaning, while in a word vector representation they are treated as different dimensions; however, because they are conceptually similar, it makes more sense to group them into the same vector dimension. Ontologies are in a good position to resolve this issue because they organize words and terms under structured concepts, group terms with similar meaning together and also maintain various semantic relations between concepts. Therefore, text can be represented on a concept level by using ontology concepts as features. Recognizing concepts in documents can further reveal semantic relations between documents (Hotho et al., 2003a), thus can facilitate further text-related tasks such as clustering, information retrieval, as well as our speech scoring task. This type of representation has been tried in several studies (e.g., Hotho et al., 2003a; Hotho et al., 2003b; Bloehdorn & Hotho, 2004).

Hotho et al. (2003a; 2003b) use ontology concepts to represent text and use the representation for document clustering. The studies employ the WordNet ontology, a general domain ontology. The experiments test three parameters of using an ontology for text representation: (1) whether concept features should be used alone or replace word features or be used together with word features; (2) word sense disambiguation strategies when using concepts; and (3) investigating the optimal level of word generalization in terms of the hierarchical structure of the ontology, i.e., how general the concepts should be. Some options of the first two parameters will be implemented and tested in our experiment design below.

The vector representation approach of text documents, either using words or concepts, can be used to measure the content similarity between essays. E-rater, for example, measures the similarity between test essays and training essays by computing the cosine similarity of their word vectors and by generating two content features based on this similarity metric. It uses multiple regression as its final scoring model, using both content features, as well as features related to other aspects of the essay, such as grammar and vocabulary usage (Burstein, 2003; Attali & Burstein, 2006). Intelligent Essay Assessor also employs cosine similarity between to-be-scored essays and training essays as basis of one content feature, and models the scoring process by normalization and regression analysis (Landauer et al., 2003). The IntelliMetric system uses a nonlinear and multidimensional modeling approach to reflect the complexity of the writing process as opposed to the general linear model (Dikli, 2006). Larkey and Croft (2003) employ Bayesian classifiers for modeling, which is a type of text categorization technique. It treats essay scoring as a text categorization task, the purpose of which is to classify essays into score categories based on content features (i.e., if the scores range from 1-4, then there are four score categories).

Zechner and Xi (2008) report on experiments related to scoring of spontaneous speech responses where content vector analysis was used as one of several features in scoring models for two different item types. They found that while these content features performed reasonably well by themselves, they were not able to increase the overall scoring model performance over a baseline that did not use content features.

This paper will use CVA as a baseline for our experiment and investigate two ontology-based approaches to enhance the content representation and improve content feature performance.

3 Data

We use data from a test for English proficiency for non-native speakers of English. Candidates are asked to provide spontaneous speech responses to four prompts, with each of the responses being one minute in length. The four prompts are all integrated prompts, meaning candidates are first given some materials to read or listen and then are asked to respond with their opinions or arguments towards the materials. The responses are scored holistically by human raters on a scale of 1 to 4, 4 being the highest score. For holistic scoring, the human raters use a speech scoring rubric as the guideline of expected performance on aspects such as fluency, pronunciation, and content for each score level.

Our data set contains 1243 speech samples in total as responses to four different prompts, obtained from 327 speakers (note that not all speakers responded to all prompts). Each response is verbatim transcribed by a human transcriber. The responses are grouped by their prompts since our experiments are prompt-specific. For responses of each prompt, we randomly split the responses into a training set (44%) and a test set (56%), making sure that response scores are distributed in a similar proportion in both training and test sets. Each response is considered as a single document here. Table 1 shows the size of the two data sets.

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 training and test data sets. The numbers in parentheses are the number of documents on score levels 1-4.

The training set is used for generating representative vectors of a prompt on different score levels, which are to be compared with test documents. The test set is primarily used to compute content features for test documents and examine performance of approaches under different experiment setups.

Besides human transcriptions of the speech files, we also obtained ASR output of the files, in order to examine performance of the proposed approaches on imperfect output, in a fully automated operational scenario where no human transcribers would be in the loop. Since the training set is used for deriving representative vectors for the four different prompts and we would like to generate accurate vectors based on human transcriptions, we do not use a separate training set for ASR data. Thus, we only obtain corresponding ASR output for the test set of each prompt.

The ASR system we use for our experiments in this paper is a state-of-the-art gender-independent continuous density Hidden Markov Model speech recognizer, trained on about 30 hours of non-native spontaneous speech. Its word error rate on the test set used here is about 12.8%.

4 Method

We employ one baseline approach for word-level features and two experimental approaches for concept-level features to examine the effect of the WordNet ontology and concept-level features on content feature correlations.

4.1 Baseline Approach: Content Vector Analysis (CVA)

We decide to use the two content features used by e-rater based on CVA analysis, called "max.cos" and "cos.w4" here (Attali & Burstein, 2006). The assumption behind this approach is that essays with similar human scores contain similar words; thus, they should share similar vector representations in CVA. For our data, this assumption is held for the spoken test documents in the same way. Moreover, we conjecture this assumption is mostly true for high score responses as opposed to low score responses, because we expect high vocabulary uniformity in high score responses and more irrelevant and more diverse vocabulary in low score responses.

Before feature computation, some preprocessing is conducted on the speech transcripts. For each prompt, we group its training set into four groups according to their score levels ("score-level documents"). Then we use the score-level documents of each prompt to generate a super vector as a representation for documents on this score level of this specific prompt. As a result, we have four score-level vectors under each prompt, generated from their training sets. While the score-level training vectors are produced using multiple documents of the same score level, vectors of test documents are generated on an individual document level. Given a test document that needs to be scored, we first convert it into the vector representation. Then we are ready to compute the two content features. Equation 1 provides the exact formula for the cosine similarity measure used in all of our methods.

$$(1)\quad \frac{\sum_{i=1}^{n} w_{t,i} * w_{sl,i}}{\sqrt{\sum_{i=1}^{n} w_{t,i}^2} * \sqrt{\sum_{i=1}^{n} w_{sl,i}^2}}$$

where n is the number of words and/or concepts in the score-level vector (from the training set documents), $w_{sl,i}$ are the word or concept weights of a score-level vector and $w_{t,i}$ are the word or concept weights of a test document (response transcription). $w_{t,i}$ are computed by term frequency and $w_{sl,i}$ are computed in the same way after concatenating documents of the same score level as one large document.

The max.cos feature. This feature measures which score level of documents the test document is most similar to in vector space by computing the cosine similarity with each score-level vector and then selecting the score level which has the largest cosine similarity to the test vector as feature value. Thus, this feature assumes integer values from 1 to 4 only.

The cos.w4 feature[2]. This feature measures content similarity between the test document and the best quality 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 an indicator of how similar the test document is to the speech content of the test takers with highest proficiency.

The two features are evaluated based on their Pearson r correlation to human assigned scores. We evaluate the features in all experiments, as a way to observe how the two features' predictiveness varies among different experiment setups. Note that since the max.cos feature assumes integer values but the cos.w4 feature is real valued, we expect correlations to be higher for cos.w4 due to this difference, all other things being equal.

4.2 Ontology-facilitated Approaches

We use two ontology-facilitated document representation approaches, which represent documents based on the WordNet ontology. The first approach matches words in a document to concepts and represents documents by vectors of concepts, whereas the second one addresses the unknown word issue by inferring their weight based on the structure of the WordNet ontology.

[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".

4.2.1 Ontology-facilitated representation approach

This representation uses concepts instead of the words as elements in the document vectors. Given a document, we map words in the document to concepts, using the synsets in WordNet. For example, *chance* and *opportunity* are different words, however they belong to the same WordNet synset (‘*opportunity.n.01*’). This concept-level representation groups words of similar meaning in the same vector dimension, thus making the vector space more succinct and semantically meaningful. The weighting scheme of concepts follows the one in the CVA approach. In this study, we focus on single words and match them to WordNet synsets; in future work, we consider matching multi-word expressions to ontologies like Wikipedia (Wikipedia, 2011). Experiments show that including words and their corresponding WordNet synsets as vector dimensions has better performance than only including WordNet synsets for text clustering tasks (Hotho et al., 2003a) and the same result also occurs in our preliminary experiments. Therefore, we include both WordNet synsets and words in the vector representation.

4.2.2 Ontology-facilitated reasoning approach

This approach is based on the ontology-facilitated representation and goes further to resolve the unknown word issue, i.e., handling words in test documents that have not been seen in the training documents.

First, test documents are converted to vectors of concepts plus words. If a concept in the test vector does not appear in the score level vector, its weight therefore is unknown, as well. We then estimate its weight based on structural information contained in the WordNet ontology. More specifically, given an unknown concept in the test document, we find the N most similar concepts to that unknown concept from the set of all concepts contained in the score level vector. We use a WordNet-based similarity estimate to measure similarity between concepts, namely the edge-based Path Similarity, which measures the length of a path from one concept to another concept in WordNet by computing the inverse of the shortest path between the two concepts (Pedersen et al., 2004). We submit that the estimated weight of the unknown concept in the test document vector should be close to the weights of its most similar concepts in the score level vector derived from the training documents. From this assumption, we propose estimating the unknown concept’s weight by averaging the weights of the N most similar concepts:

(2) $w_{unk=}(\sum_{i=1}^{N} w_i)/N$

with N denoting the number of similar concepts in a score level vector, w_i denoting the weights of these similar concepts, and w_{unk} standing for the resulting concept weight for the unknown concept in a test document.

For example, a test document may be “so radio also create a great impact on this uh people communication”. The words are matched to WordNet concepts, and we find that the concept synset ‘*impact.n.01*’ is an unknown concept to the score level 4 vector. From the dimensions of the score level 4 vector we find these three most similar concepts to the unknown concept: ‘*happening.n.01*’, ‘*event.n.01*’, and ‘*change.n.01*’. We now can average the weights of these three concepts in the score-level vector to use it as a weight estimate for the unknown concept ‘*impact.n.01*’.

We want to note that while this approach can estimate weights for test document words or concepts contained in WordNet (but not in the training vectors), it cannot handle words that are not included in WordNet at all, such as many proper names, foreign words, etc. To address the latter as well, we would have to use a much larger and more comprehensive ontology, e.g., the online encyclopedia Wikipedia.

5 Experiments and Results

We design experiments according to the above approaches. The first experiment group is the baseline system using two features employed by e-rater, max.cos and cos.w4. The second and third experiment groups implement the two ontology-facilitated approaches, respectively. We first run CVA and compare several different parameter setups to optimize them for further experiments.

5.1 Parameter Optimization in CVA Experiments

For the CVA method, we need to decide (1) which term weighting scheme to use, and (2) whether or not to use a list of stopwords to exclude common non-content words such as determiners or prepositions from consideration. We compare five commonly used term weighting schemes, each one with or without using a stoplist, based on averaged correlations with human scores across all four prompts. The best results are obtained for the weighting scheme (TF/EDL)*IDF, where TF is the frequency of a term in a document, EDL is the Euclidean document length[3], and IDF is the inverse document frequency of a term based on a collection of documents. For this scheme, as for most others, there is almost no difference between using vs. not using a stoplist and we decide to use a stoplist for our experiments based on the tradition in the field. The selected term weighting scheme is applied in the same way for both the score-level vectors as well as the test document vectors.

5.2 Experiment Groups

5.2.1 Group 1: CVA

As described above, we first convert the training sets to score level vectors and the test documents into test vectors with the TF/EDL*IDF weighting, and compute the max.cos and cos.w4 features for each test document.

5.2.2 Group 2: Ontology-facilitated Representation

We first match words in documents to WordNet concepts. There are several ways to achieve this (Hotho et al., 2003a). Given a word, it may correspond to multiple concepts in WordNet, in which each possibility is called a "sense" in WordNet, and we need to decide which sense to use.

WordNet-Sense-1. In this study we employ a simple word sense disambiguation method by using the first sense returned by WordNet. We send a word to WordNet synset search function, which returns all synstes of the word, and we select to use the first result because it is also the most frequently used sense for the word.

After obtaining the senses and concepts for the words, the training sets and test documents are converted to vectors of WordNet concepts plus words, using TF/EDL*IDF weighting, the same one used by the CVA approach. We compute the max.cos and cos.w4 features in the same way as for the baseline CVA method.

5.2.3 Group 3: Ontology-facilitated Reasoning

This approach, called here "WordNet-Reasoning", also extracts vectors of WordNet concepts plus words with the same term weighting scheme as before. For matching words to concepts, we still employ the WordNet–Sense-1 sense selection method. For unknown concepts, which appear in a test vector but not in any score level vectors, we infer their weights by using the reasoning approach proposed in section 4.2.2 with N=5 as the number of most similar concepts to the unknown concept[4], located in the WordNet hierarchy. The score level vectors are expanded by the inferred unknown concepts. When we obtain the expanded score level vectors, we compute the two content features from the vectors in the same way as before, and finally calculate feature correlations with human scores.

5.3 Results

We run the three experiment groups on human and ASR transcriptions respectively and obtain the max.cos and cos.w4 feature values of test documents in the experiments. As stated in 4.1, we compute the correlations between the two features and the human assigned scores for evaluating the approaches.

Tables 2 and 3 (next page) list correlations of the two content features with human scores under different experiment setups. Significant differences on individual prompts between correlations of the two WordNet-based methods WordNet-Sense-1 and WordNet-Reasoning and the CVA baseline are denoted with * (p<0.05) and ** (p<0.01).

[3] Given a vector of raw term frequencies $(\mathrm{rtf}_1, \mathrm{rtf}_2, \ldots, \mathrm{rtf}_\mathrm{n})$, its Euclidean length is computed in this way:

$$\sqrt{\sum_{i=1}^{n} rtf_i^{\,2}}$$

[4] We manually inspected some of the similar concepts of the unknown concepts and found the first 5 similar concepts were relevant to the unknown concepts, and thus made the decision of N=5.

Prompt	Hum, CVA	Hum, WordNet-Sense-1	Hum, Word-Net-Reasoning	ASR, CVA	ASR, Word-Net-Sense-1	ASR, Word-Net-Reasoning
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 human scores (Hum=using human transcriptions; ASR=using ASR hypotheses).

Prompt	Hum, CVA	Hum, WordNet-Sense-1	Hum, Word-Net-Reasoning	ASR, CVA	ASR, Word-Net-Sense-1	ASR, Word-Net-Reasoning
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	0.360	0.376	0.388	0.342	0.360	0.364

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

6 Discussion

6.1 Results on Human Transcriptions

On human transcriptions, Table 2 shows that the max.cos feature correlations increase, albeit not significantly, when using the method WordNet–Sense-1 on all prompts except for prompt D but decrease sometimes significantly when using the WordNet-Reasoning approach.

The cos.w4 feature correlations, on the other hand, exhibit constant increases on all four prompts when using WordNet-Sense-1 and the increase on prompt C is significant. The average correlations further increase for all prompts when using WordNet-Reasoning and the increase is significant on prompts B and C (Table 3).

6.2 Results on ASR Output

On the ASR output, for the max.cos feature, the average correlation barely changes when using the WordNet-Sense-1 method but decreases when using WordNet-Reasoning with significant decrease on prompts A and C (Table 2).

For the cos.w4 feature, however, WordNet-Sense-1 improves correlations on all four prompts with 0.018 correlation increase on average but increases are not statistically significant on a prompt level. WordNet-Reasoning does not further improve correlations much beyond the correlations of WordNet-Sense-1, with a further 0.004 increase in average correlation. Compared to CVA, though, correlations for WordNet-Reasoning are significantly higher on prompts B and C (Table 3).

6.3 Overall Discussion

Based on these observations, we find that for cos.w4, the WordNet-Sense-1 approach can improve average correlations compared to the CVA baseline on both ASR and human transcriptions. Hence, the extension of the document vectors by WordNet synsets has a positive impact on the accuracy of content scoring of the spoken responses by non-native speakers.

Again looking at the cos.w4 feature, while the WordNet Reasoning approach works well on human transcriptions to further improve correlations compared to WordNet-Sense-1, it does not consistently improve correlations on ASR output. This may indicate that WordNet-Reasoning is more sensitive to ASR errors than WordNet-Sense-1.

For the max.cos feature, the correlation of WordNet-Reasoning decreases significantly from WordNet-Sense-1 on prompts A and C for both human and ASR transcriptions; moreover, in the WordNet-Reasoning approach the max.cos correlations vary greatly on the four prompts (Table 2). We conjecture that one reason for this finding may lie in the rather small sample size of the data set, as this is an exploratory study, and the differences across prompts may be smaller when using a substantially larger data set.

Comparing the average reduction in correlation between human and ASR transcriptions, we find an absolute drop in correlations of 0.017 between the CVA baseline for the max.cos and of 0.019 for the cos.w4 feature. Looking at the WordNet-Sense-1 approach for the cos.w4 feature, the average correlation of 0.376 for human transcriptions is reduced by 0.016 to 0.360 for ASR hypotheses. Hence, we observe that the imperfect speech recognition output does not cause a major degradation for this content feature; the degradations observed are all in the range of 5% relative (the ASR word error rate on the test set is about 13%.)

Overall, the ontology-facilitated approaches are effective for the cos.w4 feature and seem to be less appropriate for the max.cos feature. We conjecture that the characteristics of the max.cos feature may be the reason for the poor performance of the ontology-facilitated approaches on this feature. To compute this feature, we need to compare a test vector with vectors for each score level, and it is assumed that these vectors are representative vectors for documents at these score levels. In reality though, while the score level 4 vector is quite a good representative for the prompt topic (highest proficiency speakers), score level vectors of less proficient speakers are less uniform and more diverse. The reason is that there are only a few ways to appropriately represent the correct topic in a good quality spoken response but there can be many different ways of generating responses that are not on topic. For example, the score level 1 vector contains vectors generated from score 1 documents, whose words are considered mostly irrelevant for the prompt. Then, given a test document, which also contains irrelevant words for the prompt but with little overlap to the level 1 score vector, the similarity between them would be very small. Thus, any ontological approach has to face this heterogeneous distribution of words in the score level vectors for responses with lower scores; any semantic generalizations are inherently more difficult compared to those on higher scoring responses. For the cos.w4 feature, in contrast, only score level 4 vectors are used, and this problem does not surface here.

Finally, we observe that average correlations of both features based on ASR hypotheses (except for WordNet-Reasoning for the max.cos feature) fall in the range of 0.32-0.37. This range is well in line with our better performing features in other dimensions of spontaneous speech responses, e.g., fluency, pronunciation, and prosody.

7 Conclusion and Future Work

In this paper, we propose using ontology-facilitated approaches for content scoring of non-native spontaneous speech due to specific merits of ontologies. Two ontology-facilitated approaches are proposed and evaluated, and their results are compared against a CVA baseline. The results indicate that the ontology approaches can improve content feature correlations in some circumstances. As a summary, concept-level features and reasoning-based approaches work well on the cos.w4 content feature where test documents are compared against a vector representing all training set documents with the highest human score.

For future work, we plan to investigate more sophisticated reasoning approaches. For this study, we use a simple averaging method to infer the concept importance based on hierarchy-inferred similarity metrics. As a next step, we plan to infer weights according to different similarity metrics and differential weighting of the N closest terms. Another avenue for future research is to employ different ontologies, for example, Wikipedia, which contains more concepts and entities than WordNet and has a structure that has grown more organically and less from first principles. Wikipedia also has a larger pool of multi-word expressions and we would like to explore how representations based on the Wikipedia ontology affects automated speech scoring performance.

References

Alwan, A., Bai, Y., Black, M., Casey, L., Gerosa, M., Heritage, M., & Wang, S. (2007). A system for technology based assessment of language and literacy in young children: The role of multiple information sources. *Proceedings of the IEEE International Workshop on Multimedia signal Processing*, Greece.

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

Balogh, J., Bernstein, J., Cheng, J., & Townshend, B. (2007). Automatic evaluation of reading accuracy: Assessing machine scores. *Proceedings of the ISCA-SLaTE-2007 Workshop, Farmington, PA, October.*

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.

Croft, W. B., Metzler, D., & Strohman, T. (2010). *Search engines: Information retrieval in practice*. Boston, MA: Addison-Wesley.

Cucchiarini, C., Strik, H., & Boves, L. (2000). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *Journal of the Acoustical Society of America*, 107(2), 989-999.

Cucchiarini, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *Journal of the Acoustical Society of America*, 111(6), 2862-2873.

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, 5(1), 1-35.

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

Franco, H., Bratt, H., Rossier, R., Gadde, V. R., Shriberg, E., Abrash, V., & Precoda, K. (2010). EduSpeak: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications. *Language Testing*, 27(3), 401-418.

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.

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.

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).*

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

Wikipedia: The free encyclopedia (2011). FL: Wikimedia Foundation, Inc. Retrieved Apr 26, 2012, from http://www.wikipedia.org

WordNet 3.0 Reference Manual. (2010). Retrieved Apr 26, 2012 from http://wordnet.princeton.edu/wordnet/documentation/

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

Zechner, K., & X. Xi (2008). Towards Automatic Scoring of a Test of Spoken Language with Heterogeneous Task Types. Proceedings of the ACL Workshop on Innovative Use of NLP for Building Educational Applications, Columbus, OH, June.