Text-to-text Semantic Similarity for Automatic Short Answer Grading

Michael Mohler and **Rada Mihalcea**
Department of Computer Science
University of North Texas
mgm0038@unt.edu, rada@cs.unt.edu

Abstract

In this paper, we explore unsupervised techniques for the task of automatic short answer grading. We compare a number of knowledge-based and corpus-based measures of text similarity, evaluate the effect of domain and size on the corpus-based measures, and also introduce a novel technique to improve the performance of the system by integrating automatic feedback from the student answers. Overall, our system significantly and consistently outperforms other unsupervised methods for short answer grading that have been proposed in the past.

1 Introduction

One of the most important aspects of the learning process is the assessment of the knowledge acquired by the learner. In a typical examination setting (e.g., an exam, assignment or quiz), this assessment implies an instructor or a grader who provides students with feedback on their answers to questions that are related to the subject matter. There are, however, certain scenarios, such as the large number of worldwide sites with limited teacher availability, or the individual or group study sessions done outside of class, in which an instructor is not available and yet students need an assessment of their knowledge of the subject. In these instances, we often have to turn to computer-assisted assessment.

While some forms of computer-assisted assessment do not require sophisticated text understanding (e.g., multiple choice or true/false questions can be easily graded by a system if the correct solution is available), there are also student answers that consist of free text which require an analysis of the text in the answer. Research to date has concentrated on two main subtasks of computer-assisted assessment: the grading of essays, which is done mainly by checking the style, grammaticality, and coherence of the essay (cf. (Higgins et al., 2004)), and the assessment of short student answers (e.g., (Leacock and Chodorow, 2003; Pulman and Sukkarieh, 2005)), which is the focus of this paper.

An automatic short answer grading system is one which automatically assigns a grade to an answer provided by a student through a comparison with one or more correct answers. It is important to note that this is different from the related task of paraphrase detection, since a requirement in student answer grading is to provide a grade on a certain scale rather than a binary yes/no decision.

In this paper, we explore and evaluate a set of unsupervised techniques for automatic short answer grading. Unlike previous work, which has either required the availability of manually crafted patterns (Sukkarieh et al., 2004; Mitchell et al., 2002), or large training data sets to bootstrap such patterns (Pulman and Sukkarieh, 2005), we attempt to devise an unsupervised method that requires no human intervention. We address the grading problem from a text similarity perspective and examine the usefulness of various text-to-text semantic similarity measures for automatically grading short student answers.

Specifically, in this paper we seek answers to the following questions. First, given a number of corpus-based and knowledge-based methods as previously proposed in the past for word and text semantic similarity, what are the measures that work best for the task of short answer grading? Second, given a corpus-based measure of similarity, what is the impact of the domain and the size of the corpus on the accuracy of the measure? Finally, can we use the student answers themselves to improve the quality of the grading system?

2 Related Work

There are a number of approaches that have been proposed in the past for automatic short answer grading. Several state-of-the-art short answer graders (Sukkarieh et al., 2004; Mitchell et al., 2002) require manually crafted patterns which, if matched, indicate that a question has been answered correctly. If an annotated corpus is avail-

able, these patterns can be supplemented by learning additional patterns semi-automatically. The Oxford-UCLES system (Sukkarieh et al., 2004) bootstraps patterns by starting with a set of keywords and synonyms and searching through windows of a text for new patterns. A later implementation of the Oxford-UCLES system (Pulman and Sukkarieh, 2005) compares several machine learning techniques, including inductive logic programming, decision tree learning, and Bayesian learning, to the earlier pattern matching approach with encouraging results.

C-Rater (Leacock and Chodorow, 2003) matches the syntactical features of a student response (subject, object, and verb) to that of a set of correct responses. The method specifically disregards the bag-of-words approach to take into account the difference between ”dog bites man” and ”man bites dog” while trying to detect changes in voice (”the man was bitten by a dog”).

Another short answer grading system, AutoTutor (Wiemer-Hastings et al., 1999), has been designed as an immersive tutoring environment with a graphical ”talking head” and speech recognition to improve the overall experience for students. AutoTutor eschews the pattern-based approach entirely in favor of a bag-of-words LSA approach (Landauer and Dumais, 1997). Later work on AutoTutor (Wiemer-Hastings et al., 2005; Malatesta et al., 2002) seeks to expand upon the original bag-of-words approach which becomes less useful as causality and word order become more important.

These methods are often supplemented with some light preprocessing, e.g., spelling correction, punctuation correction, pronoun resolution, lemmatization and tagging. Likewise, in order to facilitate their goals of providing feedback to the student more robust than a simple ”correct” or ”incorrect,” several systems break the gold-standard answers into constituent concepts that must individually be matched for the answer to be considered fully correct (Callear et al., 2001). In this way the system can determine which parts of an answer a student understands and which parts he or she is struggling with.

Automatic short answer grading is closely related to the task of text similarity. While more general than short answer grading, text similarity is essentially the problem of detecting and comparing the features of two texts. One of the earliest approaches to text similarity is the vector-space model (Salton et al., 1997) with a term frequency / inverse document frequency (*tf.idf*) weighting. This model, along with the more sophisticated LSA semantic alternative (Landauer and Dumais, 1997), has been found to work well for tasks such as information retrieval and text classification.

Another approach (Hatzivassiloglou et al., 1999) has been to use a machine learning algorithm in which features are based on combinations of simple features (e.g., a pair of nouns appear within 5 words from one another in both texts). This method also attempts to account for synonymy, word ordering, text length, and word classes.

Another line of work attempts to extrapolate text similarity from the arguably simpler problem of word similarity. (Mihalcea et al., 2006) explores the efficacy of applying WordNet-based word-to-word similarity measures (Pedersen et al., 2004) to the comparison of texts and found them generally comparable to corpus-based measures such as LSA.

An interesting study has been performed at the University of Adelaide (Lee et al., 2005), comparing simpler word and n-gram feature vectors to LSA and exploring the types of vector similarity metrics (e.g., binary vs. count vectors, Jaccard vs. cosine vs. overlap distance measure, etc.). In this case, LSA was shown to perform better than the word and n-gram vectors and performed best at around 100 dimensions with binary vectors weighted according to an entropy measure, though the difference in measures was often subtle.

SELSA (Kanejiya et al., 2003) is a system that attempts to add context to LSA by supplementing the feature vectors with some simple syntactical features, namely the part-of-speech of the previous word. Their results indicate that SELSA does not perform as well as LSA in the best case, but it has a wider threshold window than LSA in which the system can be used advantageously.

Finally, explicit semantic analysis (ESA) (Gabrilovich and Markovitch, 2007) uses Wikipedia as a source of knowledge for text similarity. It creates for each text a feature vector where each feature maps to a Wikipedia article. Their preliminary experiments indicated that ESA was able to significantly outperform LSA on some text similarity tasks.

3 Data Set

In order to evaluate the methods for short answer grading, we have created a data set of questions from introductory computer science assignments with answers provided by a class of undergraduate students. The assignments were administered as part of a Data Structures course at the University of North Texas. For each assignment, the student answers were collected via the WebCT online learning environment.

The evaluations reported in this paper are carried out on the answers submitted for three of the assignments in this class. Each assignment consisted of seven short-answer questions.[1] Thirty students were enrolled in the class and submitted answers to these assignments. Thus, the data set we work with consists of a total of 630 student answers (3 assignments x 7 questions/assignment x 30 student answers/question).

The answers were independently graded by two human judges, using an integer scale from 0 (completely incorrect) to 5 (perfect answer). Both human judges were graduate computer science students; one was the teaching assistant in the Data Structures class, while the other is one of the authors of this paper. Table 1 shows two question-answer pairs with three sample student answers each. The grades assigned by the two human judges are also included.

The evaluations are run using Pearson's correlation coefficient measured against the average of the human-assigned grades on a per-question and a per-assignment basis. In the per-question setting, every question and the corresponding student answer is considered as an independent data point in the correlation, and thus the emphasis is placed on the correctness of the grade assigned to each answer. In the per-assignment setting, each data point is an assignment-student pair created by totaling the scores given to the student for each question in the assignment. In this setting, the emphasis is placed on the overall grade a student receives for the assignment rather than on the grade received for each independent question.

The correlation between the two human judges is measured using both settings. In the per-question setting, the two annotators correlated at (r=0.6443). For the per-assignment setting, the correlation was (r=0.7228).

A deeper look into the scores given by the two annotators indicates the underlying subjectivity in grading short answer assignments. Of the 630 grades given, only 358 (56.8%) were exactly agreed upon by the annotators. Even more striking, a full 107 grades (17.0%) differed by more than one point on the five point scale, and 19 grades (3.0%) differed by 4 points or more. [2] Furthermore, on the occasions when the annotators disagreed, the same annotator gave the higher grade 79.8% of the time.

Over the course of this work, much attention was given to our choice of correlation metric. Previous work in text similarity and short-answer grading seems split on the use of Pearson's and Spearman's metric. It was not initially clear that the underlying assumptions necessary for the proper use of Pearson's metric (e.g. normal distribution, interval measurement level, linear correlation model) would be met in our experimental setup. We considered both Spearman's and several less often used metrics (e.g. Kendall's tau, Goodman-Kruskal's gamma), but in the end, we have decided to follow previous work using Pearson's so that our scores can be more easily compared.[3]

4 Automatic Short Answer Grading

Our experiments are centered around the use of measures of similarity for automatic short answer grading. In particular, we carry out three sets of experiments, seeking answers to the following three research questions.

First, *what are the measures of semantic similarity that work best for the task of short answer grading?* To answer this question, we run several comparative evaluations covering a number of knowledge-based and corpus-based measures of semantic similarity. While previous work has considered such comparisons for the related task of paraphrase identification (Mihalcea et al., 2006), to our knowledge no comprehensive evaluation has been carried out for the task of short answer grading which includes all the similarity measures proposed to date.

Second, *to what extent do the domain and the size of the data used to train the corpus-based measures of similarity influence the accuracy of the measures?* To address this question, we run a set of experiments which vary the size and domain of the corpus used to train the LSA and the ESA metrics, and we measure their effect on the accuracy of short answer grading.

Finally, *given a measure of similarity, can we integrate the answers with the highest scores and improve the accuracy of the measure?* We use a technique similar to the pseudo-relevance feedback method used in information retrieval (Rocchio, 1971) and augment the correct answer with

[1] In addition, the assignments had several programming exercises which have not been considered in any of our experiments.

[2] An example should suffice to explain this discrepancy in annotator scoring: *Question: What does a function signature include? Answer: The name of the function and the types of the parameters. Student: input parameters and return type. Scores: 1, 5.* This example suggests that the graders were not always consistent in comparing student answers to the instructor answer. Additionally, the instructor answer may be insufficient to account for correct student answers, as "return type" does seem to be a valid component of a "function signature" according to some literature on the web.

[3] Consider this an open call for discussion in the NLP community regarding the proper usage of correlation metrics with the ultimate goal of consistency within the community.

Sample questions, correct answers, and student answers	Grade
Question: What is the role of a prototype program in problem solving?	
Correct answer: To simulate the behavior of portions of the desired software product.	
Student answer 1: A prototype program is used in problem solving to collect data for the problem.	1, 2
Student answer 2: It simulates the behavior of portions of the desired software product.	5, 5
Student answer 3: To find problem and errors in a program before it is finalized.	2, 2
Question: What are the main advantages associated with object-oriented programming?	
Correct answer: Abstraction and reusability.	
Student answer 1: They make it easier to reuse and adapt previously written code and they separate complex programs into smaller, easier to understand classes.	5, 4
Student answer 2: Object oriented programming allows programmers to use an object with classes that can be changed and manipulated while not affecting the entire object at once.	1, 1
Student answer 3: Reusable components, Extensibility, Maintainability, it reduces large problems into smaller more manageable problems.	4, 4

Table 1: Two sample questions with short answers provided by students and the grades assigned by the two human judges

the student answers receiving the best score according to a similarity measure.

In all the experiments, the evaluations are run on the data set described in the previous section. The results are compared against a simple baseline that assigns a grade based on a measurement of the cosine similarity between the weighted vector-space representations of the correct answer and the candidate student answer. The Pearson correlation for this model, using an inverse document frequency derived from the British National Corpus (BNC), is r=0.3647 for the per-question evaluation and r=0.4897 for the per-assignment evaluation.

5 Text-to-text Semantic Similarity

We run our comparative evaluations using eight knowledge-based measures of semantic similarity (shortest path, Leacock & Chodorow, Lesk, Wu & Palmer, Resnik, Lin, Jiang & Conrath, Hirst & St. Onge), and two corpus-based measures (LSA and ESA). For the knowledge-based measures, we derive a text-to-text similarity metric by using the methodology proposed in (Mihalcea et al., 2006): for each open-class word in one of the input texts, we use the maximum semantic similarity that can be obtained by pairing it up with individual open-class words in the second input text. More formally, for each word *W* of part-of-speech class *C* in the instructor answer, we find $maxsim(W, C)$:

$$maxsim(W, C) = \max SIM_x(W, w_i)$$

where w_i is a word in the student answer of class *C* and the SIM_x function is one of the functions described below. All the word-to-word similarity scores obtained in this way are summed up and normalized with the length of the two input texts. We provide below a short description for each of these similarity metrics.

5.1 Knowledge-Based Measures

The **shortest path** similarity is determined as:

$$Sim_{path} = \frac{1}{length} \quad (1)$$

where $length$ is the length of the shortest path between two concepts using node-counting (including the end nodes).

The **Leacock & Chodorow** (Leacock and Chodorow, 1998) similarity is determined as:

$$Sim_{lch} = -\log \frac{length}{2 * D} \quad (2)$$

where $length$ is the length of the shortest path between two concepts using node-counting, and D is the maximum depth of the taxonomy.

The **Lesk** similarity of two concepts is defined as a function of the overlap between the corresponding definitions, as provided by a dictionary. It is based on an algorithm proposed by Lesk (1986) as a solution for word sense disambiguation.

The **Wu & Palmer** (Wu and Palmer, 1994) similarity metric measures the depth of two given concepts in the WordNet taxonomy, and the depth of the least common subsumer (LCS), and combines these figures into a similarity score:

$$Sim_{wup} = \frac{2 * depth(LCS)}{depth(concept_1) + depth(concept_2)} \quad (3)$$

The measure introduced by **Resnik** (Resnik, 1995) returns the information content (IC) of the LCS of two concepts:

$$Sim_{res} = IC(LCS) \quad (4)$$

where IC is defined as:

$$IC(c) = -\log P(c) \quad (5)$$

and $P(c)$ is the probability of encountering an instance of concept c in a large corpus.

The measure introduced by **Lin** (Lin, 1998) builds on Resnik's measure of similarity, and adds a normalization factor consisting of the information content of the two input concepts:

$$Sim_{lin} = \frac{2 * IC(LCS)}{IC(concept_1) + IC(concept_2)} \quad (6)$$

We also consider the **Jiang & Conrath** (Jiang and Conrath, 1997) measure of similarity:

$$Sim_{jnc} = \frac{1}{IC(concept_1) + IC(concept_2) - 2 * IC(LCS)} \quad (7)$$

Finally, we consider the **Hirst & St. Onge** (Hirst and St-Onge, 1998) measure of similarity, which determines the similarity strength of a pair of synsets by detecting lexical chains between the pair in a text using the WordNet hierarchy.

5.2 Corpus-Based Measures

Corpus-based measures differ from knowledge-based methods in that they do not require any encoded understanding of either the vocabulary or the grammar of a text's language. In many of the scenarios where CAA would be advantageous, robust language-specific resources (e.g. WordNet) may not be available. Thus, state-of-the-art corpus-based measures may be the only available approach to CAA in languages with scarce resources.

One corpus-based measure of semantic similarity is latent semantic analysis (LSA) proposed by Landauer (Landauer and Dumais, 1997). In LSA, term co-occurrences in a corpus are captured by means of a dimensionality reduction operated by a singular value decomposition (SVD) on the term-by-document matrix $\mathbf{T}$ representing the corpus. For the experiments reported in this section, we run the SVD operation on several corpora including the BNC (**LSA BNC**) and the entire English Wikipedia (**LSA Wikipedia**).[4]

Explicit semantic analysis (ESA) (Gabrilovich and Markovitch, 2007) is a variation on the standard vectorial model in which the dimensions of the vector are directly equivalent to abstract concepts. Each article in Wikipedia represents a concept in the ESA vector. The relatedness of a term to a concept is defined as the tf*idf score for the term within the Wikipedia article, and the relatedness between two words is the cosine of the two concept vectors in a high-dimensional space. We refer to this method as **ESA Wikipedia**.

[4]Throughout this paper, the references to the Wikipedia corpus refer to a version downloaded in September 2007.

5.3 Implementation

For the knowledge-based measures, we use the WordNet-based implementation of the word-to-word similarity metrics, as available in the WordNet::Similarity package (Patwardhan et al., 2003). For latent semantic analysis, we use the InfoMap package.[5] For ESA, we use our own implementation of the ESA algorithm as described in (Gabrilovich and Markovitch, 2006). Note that all the word similarity measures are normalized so that they fall within a 0–1 range. The normalization is done by dividing the similarity score provided by a given measure with the maximum possible score for that measure.

Table 2 shows the results obtained with each of these measures on our evaluation data set.

Measure	Correlation
Knowledge-based measures	
Shortest path	0.4413
Leacock & Chodorow	0.2231
Lesk	0.3630
Wu & Palmer	0.3366
Resnik	0.2520
Lin	0.3916
Jiang & Conrath	0.4499
Hirst & St-Onge	0.1961
Corpus-based measures	
LSA BNC	0.4071
LSA Wikipedia	0.4286
ESA Wikipedia	0.4681
Baseline	
tf*idf	0.3647

Table 2: Comparison of knowledge-based and corpus-based measures of similarity for short answer grading

6 The Role of Domain and Size

One of the key considerations when applying corpus-based techniques is the extent to which size and subject matter affect the overall performance of the system. In particular, based on the underlying processes involved, the LSA and ESA corpus-based methods are expected to be especially sensitive to changes in domain and size. Building the language models depends on the relatedness of the words in the training data which suggests that, for instance, in a computer science domain the terms "object" and "oriented" will be more closely related than in a more general text. Similarly, a large amount of training data will lead to less sparse

[5]http://infomap-nlp.sourceforge.net/

vector spaces, which in turn is expected to affect the performance of the corpus-based methods.

With this in mind, we developed two training corpora for use with the corpus-based measures that covered the computer science domain. The first corpus (**LSA slides**) consists of several online lecture notes associated with the class textbook, specifically covering topics that are used as questions in our sample. The second domain-specific corpus is a subset of Wikipedia (**LSA Wikipedia CS**) consisting of articles that contain any of the following words: computer, computing, computation, algorithm, recursive, or recursion.

The performance on the domain-specific corpora is compared with the one observed on the open-domain corpora mentioned in the previous section, namely **LSA Wikipedia** and **ESA Wikipedia**. In addition, for the purpose of running a comparison with the LSA slides corpus, we also created a random subset of the LSA Wikipedia corpus approximately matching the size of the LSA slides corpus. We refer to this corpus as **LSA Wikipedia (small)**.

Table 3 shows an overview of the various corpora used in the experiments, along with the Pearson correlation observed on our data set.

Measure - Corpus	Size	Correlation
Training on generic corpora		
LSA BNC	566.7MB	0.4071
LSA Wikipedia	1.8GB	0.4286
LSA Wikipedia (small)	0.3MB	0.3518
ESA Wikipedia	1.8GB	0.4681
Training on domain-specific corpora		
LSA Wikipedia CS	77.1MB	0.4628
LSA slides	0.3MB	0.4146
ESA Wikipedia CS	77.1MB	0.4385

Table 3: Corpus-based measures trained on corpora from different domains and of different sizes

Assuming a corpus of comparable size, we expect a measure trained on a domain-specific corpus to outperform one that relies on a generic one. Indeed, by comparing the results obtained with LSA slides to those obtained with LSA Wikipedia (small), we see that by using the in-domain computer science slides we obtain a correlation of r=0.4146, which is higher than the correlation of r=0.3518 obtained with a corpus of the same size but open-domain. The effect of the domain is even more pronounced when we compare the performance obtained with LSA Wikipedia CS (r=0.4628) with the one obtained with the full LSA Wikipedia (r=0.4286).[6] The smaller, domain-specific corpus performs better, despite the fact that the generic corpus is 23 times larger and is a superset of the smaller corpus. This suggests that for LSA the quality of the texts is vastly more important than their quantity.

When using the domain-specific subset of Wikipedia, we observe decreased performance with ESA compared to the full Wikipedia space. We suggest that for ESA the high-dimensionality of the concept space[7] is paramount, since many relations between generic words may be lost to ESA that can be detected latently using LSA.

In tandem with our exploration of the effects of domain-specific data, we also look at the effect of size on the overall performance. The main intuitive trends are there, i.e., the performance obtained with the large LSA-Wikipedia is better than the one that can be obtained with LSA Wikipedia (small). Similarly, in the domain-specific space, the LSA Wikipedia CS corpus leads to better performance than the smaller LSA slides data set. However, an analysis carried out at a finer grained scale, in which we calculate the performance obtained with LSA when trained on 5%, 10%, ..., 100% fractions of the full LSA Wikipedia corpus, does not reveal a close correlation between size and performance, which suggests that further analysis is needed to determine the exact effect of corpus size on performance.

7 Relevance Feedback based on Student Answers

The automatic grading of student answers implies a measure of similarity between the answers provided by the students and the correct answer provided by the instructor. Since we only have one correct answer, some student answers may be wrongly graded because of little or no similarity with the correct answer that we have.

To address this problem, we introduce a novel technique that feeds back from the student answers themselves in a way similar to the pseudo-relevance feedback used in information retrieval (Rocchio, 1971). In this way, the paraphrasing that is usually observed across student answers will enhance the vocabulary of the correct answer, while at the same time maintaining the correctness of the gold-standard answer.

Briefly, given a metric that provides similarity scores between the student answers and the correct answer, scores are ranked from most similar

[6] The difference was found significant using a paired t-test ($p<0.001$).

[7] In ESA, all the articles in Wikipedia are used as dimensions, which leads to about 1.75 million dimensions in the ESA Wikipedia corpus, compared to only 55,000 dimensions in the ESA Wikipedia CS corpus.

to least. The words of the top N ranked answers are then added to the gold standard answer. The remaining answers are then rescored according the the new gold standard vector. In practice, we hold the scores from the first run (i.e., with no feedback) constant for the top N highest-scoring answers, and the second-run scores for the remaining answers are multiplied by the first-run score of the Nth highest-scoring answer. In this way, we keep the original scores for the top N highest-scoring answers (and thus prevent them from becoming artificially high), and at the same time, we guarantee that none of the lower-scored answers will get a new score higher than the best answers.

The effects of relevance feedback are shown in Figure 9, which plots the Pearson correlation between automatic and human grading (Y axis) versus the number of student answers that are used for relevance feedback (X axis).

Overall, an improvement of up to 0.047 on the 0-1 Pearson scale can be obtained by using this technique, with a maximum improvement observed after about 4-6 iterations on average. After an initial number of high-scored answers, it is likely that the correctness of the answers degrades, and thus the decrease in performance observed after an initial number of iterations. Our results indicate that the LSA and WordNet similarity metrics respond more favorably to feedback than the ESA metric. It is possible that supplementing the bag-of-words in ESA (with e.g. synonyms and phrasal differences) does not drastically alter the resultant concept vector, and thus the overall effect is smaller.

8 Discussion

Our experiments show that several knowledge-based and corpus-based measures of similarity perform comparably when used for the task of short answer grading. However, since the corpus-based measures can be improved by accounting for domain and corpus size, the highest performance can be obtained with a corpus-based measure (LSA) trained on a domain-specific corpus. Further improvements were also obtained by integrating the highest-scored student answers through a relevance feedback technique.

Table 4 summarizes the results of our experiments. In addition to the per-question evaluations that were reported throughout the paper, we also report the per-assignment evaluation, which reflects a cumulative score for a student on a single assignment, as described in Section 3.

Overall, in both the per-question and per-assignment evaluations, we obtained the best performance by using an LSA measure trained on a medium size domain-specific corpus obtained from Wikipedia, with relevance feedback from the four highest-scoring student answers. This method improves significantly over the tf*idf baseline and also over the LSA trained on BNC model, which has been used extensively in previous work. The differences were found to be significant using a paired t-test ($p<0.001$).

Measure	Correlation	
	per-quest.	per-assign.
Baselines		
tf*idf	0.3647	0.4897
LSA BNC	0.4071	0.6465
Relevance Feedback based on Student Answers		
WordNet shortest path	0.4887	0.6344
LSA Wikipedia CS	**0.5099**	**0.6735**
ESA Wikipedia full	0.4893	0.6498
Annotator agreement	0.6443	0.7228

Table 4: Summary of results obtained with various similarity measures, with relevance feedback based on six student answers. We also list the tf*idf and the LSA trained on BNC baselines (no feedback), as well as the annotator agreement upper bound.

To gain further insights, we made an additional analysis where we determined the ability of our system to make a binary accept/reject decision. In this evaluation, we map the 0-5 human grading of the data set to an accept/reject annotation by using a threshold of 2.5. Every answer with a grade higher than 2.5 is labeled as "accept," while every answer below 2.5 is labeled as "reject." Next, we use our best system (LSA trained on domain-specific data with relevance feedback), and run a ten-fold cross-validation on the data set. Specifically, for each fold, the system uses the remaining nine folds to automatically identify a threshold to maximize the matching with the gold standard. The threshold identified in this way is used to automatically annotate the test fold with "accept"/"reject" labels. The ten-fold cross validation resulted in an accuracy of 92%, indicating the ability of the system to automatically make a binary accept/reject decision.

9 Conclusions

In this paper, we explored unsupervised techniques for automatic short answer grading.

We believe the paper made three important contributions. First, while there are a number of word and text similarity measures that have been proposed in the past, to our knowledge no previous work has considered a comprehensive evalu-

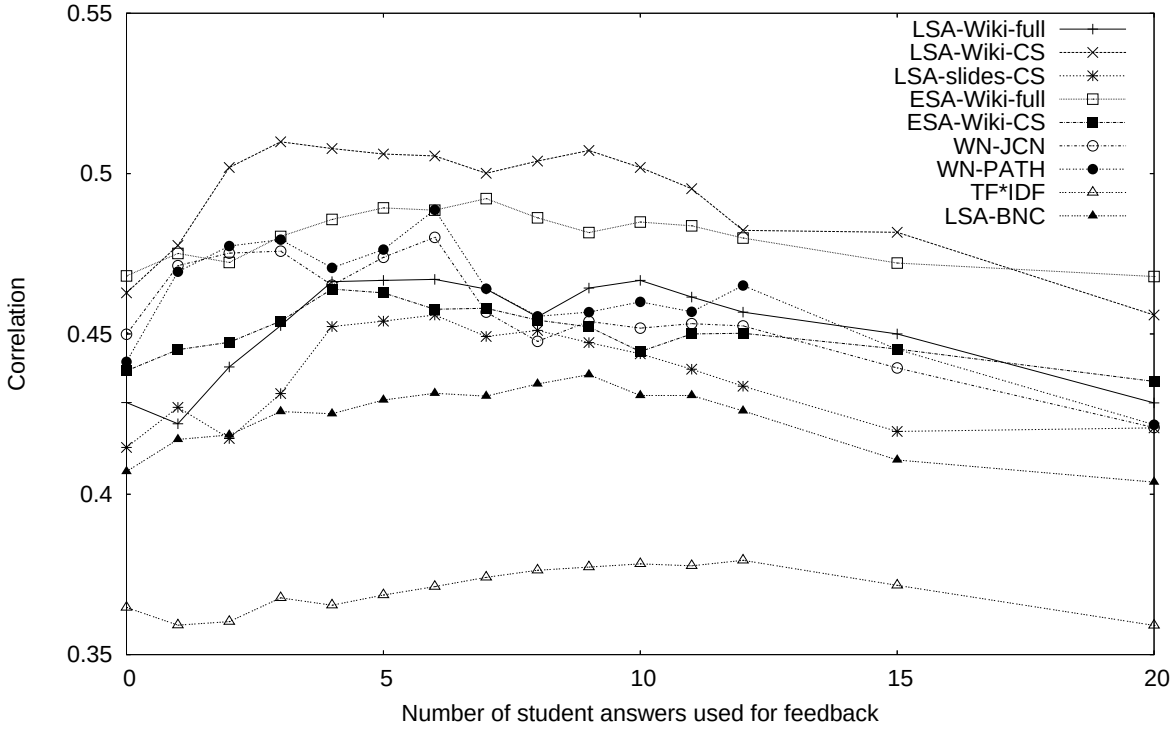

Figure 1: Effect of relevance feedback on performance

ation of all the measures for the task of short answer grading. We filled this gap by running comparative evaluations of several knowledge-based and corpus-based measures on a data set of short student answers. Our results indicate that when used in their original form, the results obtained with the best knowledge-based (WordNet shortest path and Jiang & Conrath) and corpus-based measures (LSA and ESA) have comparable performance. The benefit of the corpus-based approaches over knowledge-based approaches lies in their language independence and the relative ease in creating a large domain-sensitive corpus versus a language knowledge base (e.g., WordNet).

Second, we analysed the effect of domain and corpus size on the effectiveness of the corpus-based measures. We found that significant improvements can be obtained for the LSA measure when using a medium size domain-specific corpus built from Wikipedia. In fact, when using LSA, our results indicate that the corpus domain may be significantly more important than corpus size once a certain threshold size has been reached.

Finally, we introduced a novel technique for integrating feedback from the student answers themselves into the grading system. Using a method similar to the pseudo-relevance feedback technique used in information retrieval, we were able to improve the quality of our system by a few percentage points.

Overall, our best system consists of an LSA measure trained on a domain-specific corpus built on Wikipedia with feedback from student answers, which was found to bring a significant absolute improvement on the 0-1 Pearson scale of 0.14 over the tf*idf baseline and 0.10 over the LSA BNC model that has been used in the past.

In future work, we intend to expand our analysis of both the gold-standard answer and the student answers beyond the bag-of-words paradigm by considering basic logical features in the text (i.e., AND, OR, NOT) as well as the existence of shallow grammatical features such as predicate-argument structure(Moschitti et al., 2007) as well as semantic classes for words. Furthermore, it may be advantageous to expand upon the existing measures by applying machine learning techniques to create a hybrid decision system that would exploit the advantages of each measure.

The data set introduced in this paper, along with the human-assigned grades, can be downloaded from http://lit.csci.unt.edu/index.php/Downloads.

Acknowledgments

This work was partially supported by a National Science Foundation CAREER award #0747340. The authors are grateful to Samer Hassan for making available his implementation of the ESA algorithm.

References

D. Callear, J. Jerrams-Smith, and V. Soh. 2001. CAA of Short Non-MCQ Answers. *Proceedings of*

the 5th International Computer Assisted Assessment conference.

E. Gabrilovich and S. Markovitch. 2006. Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, Boston.

E. Gabrilovich and S. Markovitch. 2007. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 6–12.

V. Hatzivassiloglou, J. Klavans, and E. Eskin. 1999. Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning. *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

D. Higgins, J. Burstein, D. Marcu, and C. Gentile. 2004. Evaluating multiple aspects of coherence in student essays. In *Proceedings of the annual meeting of the North American Chapter of the Association for Computational Linguistics*, Boston, MA.

G. Hirst and D. St-Onge, 1998. *Lexical chains as representations of contexts for the detection and correction of malaproprisms*. The MIT Press.

J. Jiang and D. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the International Conference on Research in Computational Linguistics*, Taiwan.

D. Kanejiya, A. Kumar, and S. Prasad. 2003. Automatic evaluation of students' answers using syntactically enhanced LSA. *Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing-Volume 2*, pages 53–60.

T.K. Landauer and S.T. Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104.

C. Leacock and M. Chodorow. 1998. Combining local context and WordNet sense similarity for word sense identification. In *WordNet, An Electronic Lexical Database*. The MIT Press.

C. Leacock and M. Chodorow. 2003. C-rater: Automated Scoring of Short-Answer Questions. *Computers and the Humanities*, 37(4):389–405.

M.D. Lee, B. Pincombe, and M. Welsh. 2005. An empirical evaluation of models of text document similarity. *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, pages 1254–1259.

M.E. Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the SIGDOC Conference 1986*, Toronto, June.

D. Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, Madison, WI.

K.I. Malatesta, P. Wiemer-Hastings, and J. Robertson. 2002. Beyond the Short Answer Question with Research Methods Tutor. In *Proceedings of the Intelligent Tutoring Systems Conference*.

R. Mihalcea, C. Corley, and C. Strapparava. 2006. Corpus-based and knowledge-based approaches to text semantic similarity. In *Proceedings of the American Association for Artificial Intelligence (AAAI 2006)*, Boston.

T. Mitchell, T. Russell, P. Broomhead, and N. Aldridge. 2002. Towards robust computerised marking of free-text responses. *Proceedings of the 6th International Computer Assisted Assessment (CAA) Conference*.

Alessandro Moschitti, Silvia Quarteroni, Roberto Basili, and Suresh Manandhar. 2007. Exploiting syntactic and shallow semantic kernels for question/answer classification. In *Proceedings of the 45th Conference of the Association for Computational Linguistics*.

S. Patwardhan, S. Banerjee, and T. Pedersen. 2003. Using measures of semantic relatedness for word sense disambiguation. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, February.

T. Pedersen, S. Patwardhan, and J. Michelizzi. 2004. WordNet:: Similarity-Measuring the Relatedness of Concepts. *Proceedings of the National Conference on Artificial Intelligence*, pages 1024–1025.

S.G. Pulman and J.Z. Sukkarieh. 2005. Automatic Short Answer Marking. *ACL WS Bldg Ed Apps using NLP*.

P. Resnik. 1995. Using information content to evaluate semantic similarity. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montreal, Canada.

J. Rocchio, 1971. *Relevance feedback in information retrieval*. Prentice Hall, Ing. Englewood Cliffs, New Jersey.

G. Salton, A. Wong, and C.S. Yang. 1997. A vector space model for automatic indexing. In *Readings in Information Retrieval*, pages 273–280. Morgan Kaufmann Publishers, San Francisco, CA.

J.Z. Sukkarieh, S.G. Pulman, and N. Raikes. 2004. Auto-Marking 2: An Update on the UCLES-Oxford University research into using Computational Linguistics to Score Short, Free Text Responses. *International Association of Educational Assessment, Philadephia*.

P. Wiemer-Hastings, K. Wiemer-Hastings, and A. Graesser. 1999. Improving an intelligent tutor's comprehension of students with Latent Semantic Analysis. *Artificial Intelligence in Education*, pages 535–542.

P. Wiemer-Hastings, E. Arnott, and D. Allbritton. 2005. Initial results and mixed directions for research methods tutor. In *AIED2005 - Supplementary Proceedings of the 12th International Conference on Artificial Intelligence in Education*, Amsterdam.

Z. Wu and M. Palmer. 1994. Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, New Mexico.