On the Effectiveness of Using Syntactic and Shallow Semantic Tree Kernels for Automatic Assessment of Essays

Yllias Chali
University of Lethbridge
Lethbridge, AB, Canada
chali@cs.uleth.ca

Sadid A. Hasan
University of Lethbridge
Lethbridge, AB, Canada
hasan@cs.uleth.ca

Abstract

This paper is concerned with the problem of automatic essay grading, where the task is to grade student written essays given course materials and a set of human-graded essays as training data. Latent Semantic Analysis (LSA) has been used extensively over the years to accomplish this task. However, the major limitation of LSA is that it only retains the frequency of words by disregarding the word sequence, and the syntactic and semantic structure of texts. As a remedy, we propose the use of syntactic and shallow semantic tree kernels for grading essays. Experiments suggest that syntactic and semantic structural information can significantly improve the performance of the state-of-the-art LSA-based models for automatic essay grading.

1 Introduction and Related Work

To evaluate the content of free texts is a challenging task for humans. Automation of this process is useful when an expert evaluator is unavailable in today's Internet-based learning environment. Research to automate the assessment of free texts, such as grading student-written essays, has been carried out over the years (Kakkonen et al., 2006; Kakkonen and Sutinen, 2004; Kanejiya et al., 2003; Persing et al., 2010; Yannakoudakis et al., 2011). Some notable essay scoring systems currently available are *AutoScore* by American Institutes for Research (AIR), *Bookette* by CTB/McGraw-Hill, *Project Essay Grade* by Measurement, Inc. and *Intelligent Essay Assessor* by Pearson Knowledge Technologies. The approaches such as Project Essay Grade and e-rater were solely based on some simple surface features that took essay-length, number of commas etc. into consideration (Page and Petersen, 1995; Powers et al., 2000). The major drawback of these systems is that they ignore the creativity factor by only dealing with the simple measures. To overcome this limitation, recent researches tend to focus on understanding the inner meaning of the texts. Latent Semantic Analysis (LSA) (Landauer et al., 1998; Deerwester et al., 1990) has been shown to fit well in addressing this task (Kakkonen et al., 2006; Kakkonen and Sutinen, 2004; Lintean et al., 2010; Kanejiya et al., 2003).

LSA uses a sophisticated approach to decode the inherent relationships between a context (typically a sentence, a paragraph or a document) and the words that they contain. This approach is based on Bag-Of-Words (BOW) assumption that uses the frequency of occurrence of each word in the context to construct a word-by-context co-occurrence matrix (Kanejiya et al., 2003). The major limitation of LSA is that it only retains the frequency of the words and does not take into account the sequence of them (word ordering). It ignores the syntactic and semantic structure of the context and thus, cannot distinguish between "The police shot the gunman" and "The gunman shot the police". Traditionally, information extraction techniques are based on the BOW approach augmented by language modeling. But when the task like *automated essay grading* requires the evaluation of more complex syntactic and semantic structures, the approaches based on only BOW are often inadequate to perform fine-level textual analysis. For example, in the basic LSA model for automated essay grading, a student essay can obtain a good grade by having a very small number of highly representative words that correlates the golden essays. This also means that the repetition of important terms without having any syntactic/semantic appropriateness can lead to a overstated grade (Jorge-Botana et al., 2010).

Several improvements on BOW have been shown by the use of dependency trees and syntac-

International Joint Conference on Natural Language Processing, pages 767–773,
Nagoya, Japan, 14-18 October 2013.

tic parse trees over the years (Hirao et al., 2004; Punyakanok et al., 2004; Kim and Kim, 2010). Kakkonen et al. (2006) used an enhanced LSA approach by incorporating parts-of-speech (POS) information to improve the performance of the basic LSA model for automatic essay grading. The augmentation of POS information into the basic LSA model enabled it to exploit a sufficient amount of local information about internal relations among the words. In this manner, the enhanced LSA model could disambiguate the meaning between the words having the same base forms but different POS tags. Kanejiya et al. (2003) proposed a similar model called *Syntactically Enhanced LSA* by considering a word along with its syntactic neighborhood (obtained from the part-of-speech tag of its preceding word). Wiemer-Hastings and Zipitria (2001) showed that a sentence comparison metric that combines structure-derived information with vector-based semantics has a better correlation to human judgements than the LSA model alone. This motivates us to propose the use of syntactic and semantic structural information (by means of syntactic and shallow semantic tree kernels) with a LSA-based model to automatically grade essays. The effectiveness of using various text-to- text semantic similarity measures, and dependency graph alignment techniques have been also shown to improve upon the BOW approaches for a similar task of short answer grading (Mohler et al., 2011; Mohler and Mihalcea, 2009).

The importance of syntactic and semantic features in finding textual similarity is described by Moschitti et al. (2007), and Moschitti and Basili (2006). An effective way to integrate syntactic and semantic structures in different applications is the use of *tree kernel* functions (Collins and Duffy, 2001), which has been successfully applied to other Natural Language Processing (NLP) tasks such as question classification (Moschitti and Basili, 2006). In this paper, we use the tree kernel functions and to the best of our knowledge, no other study has used tree kernel functions before to encode syntactic/semantic information for more complex tasks such as computing the relatedness between the contexts for automatic essay grading. Our experiments on an occupational therapy dataset show that the addition of syntactic and semantic information can improve the performance of the BOW-based and POS enhanced state-of-the-art LSA models significantly.

2 LSA Model for Essay Grading

LSA can determine the similarity of the meaning of words and the context based on word co-occurrence information (Kakkonen et al., 2006). Our grading model is most closely related to the approach described in Kakkonen and Sutinen (2004) where the experiments were conducted in the Finnish language. However, in this work, we experiment with the essays and course materials written in the English language. The main idea is based on the assumption that a student's knowledge is largely dependent on learning the course content; therefore, the student's knowledge can be computed as the degree of semantic similarity between the essay and the given course materials. An essay will get a higher grade if it closely matches with the course content.

The grading process includes three major steps. In the first step, we build a semantic space from the given course materials by constructing a word-by-context matrix (WCM). Here we use different local and global weighting functions to build several LSA models (for baseline selection). In the next step, a set of pre-scored (human-graded) essays are transformed into a query-vector form similar to each vector in the WCM and then their similarity with the semantic space is computed in order to define the threshold values for each grade category. The similarity score for each essay is calculated by using the traditional cosine similarity measure. In the last step, the student-written to-be-graded essays are transformed into the query-vector forms and compared to the semantic space in a similar way. The threshold values for the grade categories are examined to specify which essay belongs to which grade category.

As discussed previously, the basic LSA model for automatic essay grading lacks sensitivity to the context in which the words appear since it is solely based on the BOW assumption. It ignores the internal structure of the sentences and does not consider word orders. Our aim in this paper is to propose a similarity measure in which syntactic and/or semantic information can be added to enhance the basic LSA model by encoding the relational information between the words in sentences. We claim that for a complex task like evaluating student-written essays, where the relatedness between the sentences of an essay and the given course materials is an important factor, our grading model would perform more effectively if we could incorporate the syntactic and semantic information with the standard cosine measure (i.e. done in basic LSA) while calculating the similarity between sentences. In the next sections, we describe how we can encode syntactic and semantic structures in calculating the similarity between sentences.

3 Syntactic Similarity Measure (SYN)

Inspired by the potential significance of using syntactic measures for finding similar texts, we get a strong motivation to use it as a similarity measure in essay grading framework. The first step to calculate the syntactic similarity between two sentences is to parse the corresponding sentences into syntactic trees using the Charniak parser (Charniak, 1999). Once we build the syntactic trees, our next task is to measure the similarity between the trees. For this, every tree T is represented by an m dimensional vector $v(T) = (v_1(T), v_2(T), \cdots v_m(T))$, where the i-th element $v_i(T)$ is the number of occurrences of the i-th tree fragment in tree T (Moschitti et al., 2007). The tree kernel of two trees T_1 and T_2 is actually the inner product of $v(T_1)$ and $v(T_2)$ (Collins and Duffy, 2001), which computes the number of common subtrees between two trees to provide the similarity score between a pair of sentences. Each course material sentence contributes a score to the essay sentences. The average syntactic similarity scores of the essay sentences are combined to get an overall similarity score for an essay with respect to the course material sentences.

4 Semantic Similarity Measure (SEM)

Shallow semantic representations can prevent the weakness of cosine similarity based models (Moschitti et al., 2007). Since the textual similarity between a pair of sentences relies on a deep understanding of the semantics of both, applying semantic similarity measurement in our essay grading framework is another noticeable contribution of this paper. To calculate the semantic similarity between two sentences, we first parse the corresponding sentences semantically using the Semantic Role Labeling (SRL) system, ASSERT[1]. We represent the annotated sentences using tree structures called semantic trees (ST). In the tree kernel method (Section 3), common substructures cannot be composed of a node with only some of its children. Moschitti et al. (2007) solved this problem by designing the Shallow Semantic Tree Kernel (SSTK) which allows to match portions of a ST. The SSTK function yields the similarity score between a pair of sentences based on their semantic structures. An overall semantic similarity score for each essay is obtained similarly as the syntactic measure.

5 Experiments and Evaluation

5.1 Data

We use a dataset obtained from an occupational therapy course where 3 journal articles are provided as the course materials. The students are asked to answer an essay-type question. The dataset contains 91 student-written essays, which are graded by a professor[2]. The length of the essays varied from 180 to 775 characters. We use 3-fold cross-validation for our experiments.

5.2 System Settings

Initially, we split the course materials into 64 paragraphs and built the word-by-paragraph matrix by treating the paragraphs as contexts. Our preliminary experiments suggested that this scheme shows worse performance than that of using individual sentences as the contexts. So, we tokenized the course materials (journal articles) into 741 sentences and built the word-by-sentence matrix. We do not perform word stemming for our experiments. We use a stop word list of 429 words to remove any occurrence of them from the datasets. In this work, C++ and Perl are used as the programming languages to implement the LSA models and encode the syntactic and shallow semantic structures. The GNU Scientific Library (GSL[3]) software package is used to perform the SVD calculations in LSA. During the dimensionality reduction step of LSA, we have experimented with different dimensions of the semantic space. Finally, we kept 100 as the number of dimensions since we got better results using this value. We experiment with six variations of the LSA model based on different local and global weighting functions according to Chali and Hasan (2012). The best performing LSA model is used as the baseline for comparison purposes.

[1] Available at http://cemantix.org/assert

[2] Each essay is graded on a scale from 0 to 6.

[3] http://www.gnu.org/software/gsl/

5.2.1 Variations of the LSA Model

Inspired by the work of Jorge-Botana et al. (2010), we experiment with different local and global weighting functions applied to the WCM. The main idea is to transform the raw frequency cell x_{ij} of the WCM into the product of a local term weight l_{ij}, and a global term weight g_j. Given the term/document frequency matrix (WCM), a weighting algorithm is applied to each entry that has three components to makeup the new weighted value in the term/document matrix. This looks as: $w_{ij} = l_{ij} * g_j * N_j$, where w_{ij} is the weighted value for the i^{th} term in the j^{th} context, l_{ij} is the local weight for term i in the context j, g_j is the global weight for the term i across all contexts in the collection, and N_j is the normalization factor for context j.

Local Weighting: We use two local weighting methods in this work: *1) Logarithmic:* $\log(1 + f_{ij})$, and *2) Term Frequency (TF):* f_{ij}, where f_{ij} is the number of times (frequency) the term i appears in the context j.

Global Weighting: We experiment with three global weighting methods: *1) Entropy:* $1 + \left(\frac{\sum_j (p_{ij} \log(p_{ij}))}{\log(n)}\right)$, *2) Inverse Document Frequency (IDF):* $\log\left(\frac{n}{df_i}\right) + 1$, and *3) Global Frequency/Inverse Document Frequency (GF/IDF):* $\frac{\sum_j f_{ij}}{df_i}$, where $p_{ij} = \frac{f_{ij}}{\sum_j f_{ij}}$, n is the number of documents in our word by context matrix, and df_i is the number of contexts in which the term i is present.

Different Models: By combining the different local and global weighting schemes, we build the following six different LSA models: **1) LE:** logarithmic local weighting and entropy-based global weighting, **2) LI:** logarithmic local weighting and IDF-based global weighting, **3) LG:** logarithmic local weighting and GF/IDF-based global weighting, **4) TE:** TF-based local weighting and entropy-based global weighting, **5) TI:** TF-based local weighting and IDF-based global weighting, and **6) TG:** TF-based local weighting and GF/IDF-based global weighting.

5.2.2 Systems for Evaluation

To study the impact of syntactic and semantic representation introduced earlier (in Section 3 and Section 4) for the essay grading task, we build six systems as defined below:

(1) Baseline: Our baseline is the best performing LSA model among the six variations (discussed in Section 5.2.1) that uses the standard cosine similarity measure based on BOW assumption and does not consider syntactic/semantic information.

(2) SYN: This system measures the similarity between the sentences using the *syntactic tree* and the *general tree kernel* function defined in Section 3.

(3) SEM: This system measures the similarity between the sentences using the *shallow semantic tree* and the *shallow semantic tree kernel* function defined in Section 4.

(4) LSA+SYN: This system measures the similarity between the sentences using both standard cosine similarity measure and the syntactic tree kernel.

(5) LSA+SEM: This system measures the similarity between the sentences using both standard cosine similarity measure and the shallow semantic tree kernel.

(6) LSA+SYN+SEM: This system measures the similarity between the sentences using standard cosine similarity measure, syntactic tree kernel, and shallow semantic tree kernel.

We use an equally weighted linear combination by summing the similarity scores obtained by **LSA**, **SYN** and **SEM** (when multiple similarity measures are used) as we believe that the word distribution, syntactic and semantic similarity between a pair of texts are all equally important. The average value of the similarity scores of the representative essays (with comparison to the course materials) of a certain grade category is considered as the threshold for that particular grade. For example, if we have five pre-scored essays of grade 6, we obtain five similarity scores corresponding to the course materials. The average of these scores are considered as the minimum score (threshold) that should be obtained by a non-graded student-written essay in order to assign it the grade 6. For a more robust evaluation, we also implement a state-of-the-art part-of-speech (POS) enhanced LSA model (**POS+LSA**) for essay grading according to Kakkonen et al. (2006) by considering the POS tag of the current word.

5.3 Evaluation Results

In Table 1, we present the results of our baseline selection step. The first column stands for the weighting model used ("N" denotes no weighting method applied). The "Correlation" column presents the Spearman rank correlation between the scores given by the professor and the systems. The "Accuracy" column stands for the proportion of the cases where the professor and the system have assigned the same grade whereas the next column shows the percentage of essays where the system-assigned grade is at most one point away or exactly the same as the professor. From these results, we can see that the performance of the systems varied (having correlation from 0.32 to 0.68) with respect to the weighting scheme applied. We observe that the combination of the logarithmic local weighting with the entropy-based global weighting scheme performs the best for our dataset. Hence, we use this model as our baseline system.

In Table 2, we present the results of different systems. The columns denote the same meaning as Table 1. We can see that for the **SYN** system, the correlation is decreased by 7.93% from the baseline and 12.69% from the **POS+LSA** system. The **SEM** system improves the correlation over the baseline system by 2.94%, but decreases by 1.42% from the **POS+LSA** system. The **LSA+SYN** system improves the correlation over the baseline system by 7.35% and over the **POS+LSA** system by 2.81% whereas the **LSA+SEM** system improves the correlation by 11.76%, and 7.04% respectively. Lastly, the **LSA+SYN+SEM** system improves the correlation over the baseline system by 10.29% and over the **POS+LSA** system by 5.63%. Analysis of these results reveals that the proposed systems (that encode the syntactic and/or semantic information with the basic LSA model) considerablely outperform both the standard cosine similarity based and the state-of-the-art POS enhanced LSA approaches. The results also denote that encoding the syntactic and/or semantic information on top of the standard cosine similarity measure often outperform the systems that consider only syntactic and/or semantic information.

Statistical Significance: We use Student's t-test to compute whether the differences between the correlations of different systems are statistically significant. For this computation, we have one measurement variable, "correlation", and one nominal variable, "system". We had three runs and the observations were the set of correlations for each of the systems in consideration. We find that the differences between the correlations are statistically significant at $p < 0.05$ except for the differences between the **SEM** system and the **POS+LSA** system, and between the **LSA+SYN+SEM** system and the **LSA+SEM** system. We also compute the statistical significance of the correlations themselves. In Table 1, the reported correlations are statistically significant ($p < 0.05$) except for "TE" and "N" models. The correlations reported in Table 2 are statistically significant ($p < 0.05$).

Model	Corr.	Accuracy (%)	Close (%)
LE	0.68	40.2	73.1
LI	0.49	27.1	51.8
LG	0.40	21.3	42.2
TE	0.34	19.2	36.4
TI	0.52	32.6	58.6
TG	0.38	20.4	38.9
N	0.32	17.8	32.9

Table 1: Variations of LSA model

System	Corr.	Accuracy (%)	Close (%)
Baseline	0.68	40.2	73.1
POS+LSA	0.71	42.6	70.8
SYN	0.63	34.8	60.1
SEM	0.70	41.5	76.2
LSA+SYN	0.73	43.2	78.1
LSA+SEM	0.76	48.3	82.5
LSA+SYN+SEM	0.75	46.7	79.6

Table 2: Evaluation results

5.4 Discussion

5.4.1 Is Thresholding Adequate?

Our experiments showed that the formation of the thresholds were adequate as we could obtain different thresholds for different grade categories. However, in a few cases, the difference between two subsequent thresholds was found to be small. This might be because the grades were not evenly distributed among the given human-graded corpus. Ideally it is desirable to have the representative training essays across the spectrum of possible grades to set the thresholds on by using the SVD generated from the training materials. We also believe that the use of a larger dataset while defining the thresholds might improve the overall performance. Our further experiments (shown in the next subsection) support this claim. The length of the essays is another issue since longer essays tend to capture more information in their representative vectors which provides the scope for a better similarity matching with the semantic space.

5.4.2 Can We Automate Data Generation?

To experiment with an LSA-based model we require a number of student-written essays. It is often hard to collect a huge number of raw student-written essays and process them into the machine-readable format. To reduce the human intervention involved in producing a large amount of training data, we propose to automate this process by using the ROUGE (Lin, 2004) toolkit. We assume each individual sentence of the course material as the candidate extract sentence and calculate its ROUGE similarity scores with the corresponding golden essay. Thus an average ROUGE score is assigned to each sentence of the course content. We choose the top 50% sentences based on ROUGE scores to have the label +1 (candidate essay sentences) and the rest to have the label -1 (non-essay sentences), and thus, we generate essays up to a predefined word limit considering different levels of expertise of the students. The sentences having the label +1 are further sorted in descending order of their assigned scores. A collection of sentences (upto length 775 characters) having the highest scores are considered to have the grade 6, the next collection of sentences to grade 5 and so on. In this manner, we have generated 216 essays from the given course materials. We have used 20 golden essays in this experiment. We treated the essays that got the full score of 6 as the golden essays. The automatically generated essays appeared to be similar in content to that of the original student-written essays.

We run further experiments using the automatically generated dataset in order to make sure that the proposed methods are useful for the essay grading task. For this purpose, we build a corpus containing 147 essays (that include both human-written and automatic essays), where the grade categories are evenly distributed. We use 3-fold cross-validation for our experiments. In Table 3, we present the results of different systems. A relative comparison of these results with the results of Table 2 yields that there is a marginal improvement in the overall performance of all the systems except for the **LSA+SYN** system. This phenomenon suggests that the even distribution of the grade categories in a larger corpus of essays is useful in general to achieve better grading performance. The results also reveal the effectiveness of our proposed method for automatic training data generation. The differences between the correlations are statistically significant at $p < 0.05$ (using Student's t-test) except for the differences between the **LSA+SYN** system and the baseline, and between the **LSA+SYN+SEM** system and the **LSA+SEM** system. The reported correlations are also found to be statistically significant ($p < 0.05$).

System	Corr.	Accuracy (%)	Close (%)
Baseline	0.71	42.6	75.4
POS+LSA	0.73	45.2	72.5
SYN	0.65	35.2	63.5
SEM	0.75	48.5	79.7
LSA+SYN	0.72	42.8	77.5
LSA+SEM	0.80	52.3	84.2
LSA+SYN+SEM	0.78	50.1	81.6

Table 3: Evaluation results (second corpus)

6 Conclusion and Future Work

We proposed to encode the syntactic and semantic information for measuring sentence relationships to automatically grade student-written essays and demonstrated that adding syntactic and/or semantic information on top of the standard cosine measure improves the performance over the BOW based and state-of-the-art POS enhanced LSA models. To the best of our knowledge, no other study has used syntactic and shallow semantic tree kernels for the task of automatic essay grading to improve the basic LSA model's performance. Our approach to automate the data generation process is also unique and novel in this problem domain. Experimental results revealed the effectiveness of the proposed approach. Our experiments also suggested that the overall syntactic/semantic similarity between a pair of texts can be effectively captured using the aggregated tree kernel scores of all possible sentence pairs. In the future, we plan to focus on other important metrics in terms of creativity, novelty, etc. for the essay grading task which we believe would further enhance the overall grading performance given that the major limitation of the basic LSA model is overcome.

Acknowledgments

The research reported in this paper was supported by the Mitacs-Accelerate internship program, the Natural Sciences and Engineering Research Council (NSERC) of Canada – discovery grant and the University of Lethbridge. The authors are grateful to Colin Layfield and Laurence Meadows for their assistance.

References

Y. Chali and S. A. Hasan. 2012. Automatically Assessing Free Texts. In *Proceedings of the Workshop on Speech and Language Processing Tools in Education*, pages 9–16, Mumbai, India. COLING 2012.

E. Charniak. 1999. A Maximum-Entropy-Inspired Parser. In *Technical Report CS-99-12*, Brown University, Computer Science Department.

M. Collins and N. Duffy. 2001. Convolution Kernels for Natural Language. In *Proceedings of Neural Information Processing Systems*, pages 625–632, Vancouver, Canada.

S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

T. Hirao, J. Suzuki, H. Isozaki, and E. Maeda. 2004. Dependency-based Sentence Alignment for Multiple Document Summarization. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 446–452, Geneva, Switzerland.

G. Jorge-Botana, J. A. Leon, R. Olmos, and I. Escudero. 2010. Latent Semantic Analysis Parameters for Essay Evaluation using Small-Scale Corpora. *Journal of Quantitative Linguistics*, 17(1):1–29.

T. Kakkonen and E. Sutinen. 2004. Automatic Assessment of the Content of Essays Based on Course Materials. In *Proceedings of the 2nd IEEE International Conference on Information Technology: Research and Education*, pages 126–130.

T. Kakkonen, N. Myller, and E. Sutinen. 2006. Applying Part-Of-Speech Enhanced LSA to Automatic Essay Grading. In *Proceedings of the 4th IEEE International Conference on Information Technology: Research and Education (ITRE 2006)*.

D. Kanejiya, A. Kumar, and S. Prasad. 2003. Automatic Evaluation of Students' Answers using Syntactically Enhanced LSA. In *Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing - Volume 2*, pages 53–60. ACL.

Y. Kim and Y. Kim. 2010. An Autonomous Assessment System based on Combined Latent Semantic Kernels. *Expert Systems with Applications*, 37(4):3219–3228.

T. Landauer, P. Foltz, and D. Laham. 1998. An Introduction to Latent Semantic Analysis. *Discourse Processes*, 25(2):259–284.

C. Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of Association for Computational Linguistics*, pages 74–81, Barcelona, Spain.

M. C. Lintean, C. Moldovan, V. Rus, and D. S. McNamara. 2010. The Role of Local and Global Weighting in Assessing the Semantic Similarity of Texts Using Latent Semantic Analysis. In *FLAIRS Conference*.

M. Mohler and R. Mihalcea. 2009. Text-to-Text Semantic Similarity for Automatic Short Answer Grading. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 567–575. ACL.

M. Mohler, R. Bunescu, and R. Mihalcea. 2011. Learning to Grade Short Answer Questions using Semantic Similarity Measures and Dependency Graph Alignments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, pages 752–762. ACL.

A. Moschitti and R. Basili. 2006. A Tree Kernel Approach to Question and Answer Classification in Question Answering Systems. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genoa, Italy.

A. Moschitti, S. Quarteroni, R. Basili, and S. Manandhar. 2007. Exploiting Syntactic and Shallow Semantic Kernels for Question/Answer Classificaion. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007)*, pages 776–783, Prague, Czech Republic.

E. B. Page and N. S. Petersen. 1995. The Computer Moves into Essay Grading: Updating the Ancient Test. *Phi Delta Kappan*, 76(7).

I. Persing, A. Davis, and V. Ng. 2010. Modeling Organization in Student Essays. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 229–239. ACL.

D. E. Powers, J. Burstein, M. Chodorow, M. E. Fowles, and K. Kukich. 2000. Comparing the Validity of Automated and Human Essay Scoring. *(GRE No. 98-08a, ETS RR-00-10). Princeton, NJ: Educational Testing Service.*

V. Punyakanok, D. Roth, and W. Yih. 2004. Mapping Dependencies Trees: An Application to Question Answering. In *Proceedings of AI & Math*, Florida, USA.

P. Wiemer-Hastings and I. Zipitria. 2001. Rules for Syntax, Vectors for Semantics. In *Proceedings of the Twenty-third Annual Conference of the Cognitive Science Society*, pages 1112–1117. Erlbaum.

H. Yannakoudakis, T. Briscoe, and B. Medlock. 2011. A New Dataset and Method for Automatically Grading ESOL Texts. In *Proceedings of the 49th ACL-HLT*, pages 180–189. ACL.