Automatically Assessing Free Texts

Yllias Chali *Sadid A. Hasan*
University of Lethbridge, Lethbridge, AB, Canada
chali@cs.uleth.ca, hasan@cs.uleth.ca

ABSTRACT
Evaluation of the content of free texts is a challenging task for humans. Automation of this process is largely useful in order to reduce human related errors. We consider one instance of the "free texts" assessment problems; automatic essay grading where the task is to grade student written essays automatically given course materials and a set of human-graded essays as training data. We use a Latent Semantic Analysis (LSA)-based methodology to accomplish this task. We experiment on a dataset obtained from an occupational therapy course and report the results. We also discuss our findings, analyze different problem areas and explain the potential solutions.

KEYWORDS: Free texts, essay grading, latent semantic analysis, syntactic tree kernel, shallow semantic tree kernel.

Proceedings of the Workshop on Speech and Language Processing Tools in Education, pages 9–16,
COLING 2012, Mumbai, December 2012.

1 Introduction

The problem of assessing free texts involves understanding the inner meaning of the free texts. Automation of free text assessment is necessary specially when an expert evaluator is unavailable in today's Internet-based learning environment. This is also useful to reduce human related errors such as "rater effects" (Rudner, 1992). Research to automate the assessment of free texts, such as grading student-written essays, has been carried out over the years. The earlier approaches such as Project Essay Grade (PEG) (Page and Petersen, 1995) and e-rater (Powers et al., 2000) were solely based on some simple surface features that took essay-length, number of commas etc. into consideration whereas recent research has tended 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 previously (Kakkonen et al., 2006; Kakkonen and Sutinen, 2004; Briscoe et al., 2010). 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. The main idea behind the LSA is to measure the semantic similarities to be found between two texts from words contained within. In this paper, we use LSA to automatically grade student-written essays. We experiment with different local and global weighting functions[1]. Experiments on an occupational therapy dataset show that the performance of the LSA varies with respect to the weighting function used. In the next sections, we present an overview of LSA, describe our approach, and present the evaluation results. We then discuss various problem areas related to the evaluation framework and explain potential solutions. Finally, we conclude the paper.

2 Overview of LSA

LSA, that has been used successfully in various NLP tasks (Cederberg and Widdows, 2003; Clodfelder, 2003; Kanejiya et al., 2003; Pino and Eskenazi, 2009), can determine the similarity of the meaning of words and the context based on word co-occurrence information (Kakkonen et al., 2006). In the first phase of LSA, a word-by-context (WCM) matrix is constructed that represents the number of times each distinct word appears in each context. Next, weighting may be applied to the values contained in this matrix in relation to their frequency in order to better represent the importance of a word. The main idea of using a weighting function is to give higher values to the words that are more important for the content and lower values otherwise (Kakkonen and Sutinen, 2004). The next phase is called the dimensionality reduction step. In this phase, the dimension of the WCM is reduced by applying Singular Value Decomposition (SVD) and then reducing the number of singular values in SVD. This is done in order to try and draw out underlying latent semantic similarities between texts in the decomposition when comparison operators are used. This step also enables words that are used in a similar fashion, but not necessarily in the same documents, to be viewed as having a similar role (synonymy) in the texts, thus, enhancing their similarity scores. By reducing the dimensions, LSA can enhance the score of two similar documents whilst decreasing the score of non similar documents. Thus the process makes the context and the words more dependent to each other by reducing the inherent noise of the data set (Jorge-Botana et al., 2010).

3 Our Approach

Our approach 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

[1] An estimation to calculate the representativeness of a word in a document.

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. 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.

4 Experiments and Evaluation

4.1 System Description

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:* $\log\Big(1 + f_{ij}\Big)$, 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.

4.2 Implementation

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. For our experiments, we randomly choose 61 pre-scored essays to build the threshold values for different grade categories, and the rest of the essays are used as the test data. 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. The GNU Scientific Library (GSL[3]) software package is used to perform the SVD calculations. During the dimensionality reduction step, 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.

4.3 Results and Discussion

In Table 1, we present the results of our experiments. The first column stands for the weighting model used ("N" denotes no weighting method applied, which acts as a baseline for this work). 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 see that the combination of the logarithmic local weighting with the entropy-based global weighting scheme performs the best for our dataset. However, the reason behind lower correlations of all the LSA models might be that the threshold values for the grade categories became largely dependent on the training essays and the course materials. This is because the grades were not evenly distributed among the given human-graded corpus (see Table 2). 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 believe that the use of a larger dataset while defining the thresholds might improve the LSA model's performance. 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.

Weighting Model	Correlation	Accurate (%)	Accurate or one point away (%)
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: Results

[2]Each essay is graded on a scale from 0 to 6.
[3]http://www.gnu.org/software/gsl/

Grade	Distribution (%)
0	1.10
1	1.10
2	1.10
3	12.08
4	25.27
5	24.17
6	35.16

Table 2: Grade distribution

5 Analyses and Solutions

5.1 Automating the Evaluation

The performance of the LSA models can be verified by measuring their correlation with the human-graded essays (as shown in Section 4.3). To omit the human intervention associated with this method, we can introduce an automatic evaluation module that uses syntactic and/or shallow semantic tree kernels to measure the textual similarity between the student-written essays and the given course materials. The basic LSA model that uses cosine similarity measure has one problem in automatic grading of academic essays. In this method, a student essay can obtain a good grade by having a very small number of highly representative terms 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). So, we can check the LSA model's performance by measuring syntactic/semantic similarity of the student-written essays corresponding to the course materials. Syntactic and semantic features have been used successfully in various NLP tasks (Zhang and Lee, 2003; Moschitti et al., 2007; Moschitti and Basili, 2006). Based on some preliminary case-by-case analysis, we find the automatic evaluation model to be promising.

Syntactic Tree Kernel: Given the sentences, we can first parse them into syntactic trees using a parser like (Charniak, 1999) and then, calculate the similarity between the two trees using the *tree kernel* (Collins and Duffy, 2001). Once we build the trees (syntactic or semantic), our next task is to measure the similarity between the trees. For this, every tree T is represented by an m dimensional vector $v(T) = \big(v_1(T), v_2(T), \cdots v_m(T)\big)$, where the i-th element $v_i(T)$ is the number of occurrences of the i-th tree fragment in tree T. The tree kernel of two trees T_1 and T_2 is actually the inner product of $v(T_1)$ and $v(T_2)$: $TK(T_1, T_2) = v(T_1).v(T_2)$. TK is the similarity value (tree kernel) between a pair of sentences based on their syntactic structure.

Shallow Semantic Tree Kernel (SSTK): To calculate the semantic similarity between two sentences, we first parse the corresponding sentences semantically using the Semantic Role Labeling (SRL) (Moschitti et al., 2007; Kingsbury and Palmer, 2002; Hacioglu et al., 2003) system, ASSERT[4]. We represent the annotated sentences using tree structures called semantic trees (ST). The similarity between the two STs is computed using the shallow semantic tree kernel (SSTK) (Moschitti et al., 2007). This is the semantic similarity score between a pair of sentences based on their semantic structures.

5.2 Automating Data Generation

To experiment with the 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

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

Essays (Score 6)	Example
Automatic	Since it seemed unlikely that Hans will be able to completely return to his former life structure, the following goals were established for his habituation: To modify Hans' habit patterns (i.e., identify new leisure activities to build into his schedule, especially on the weekend.), To enable Hans to acquire a new role (i.e., the role of a volunteer), To assist Hans to modify some of his roles (e.g., being a spectator or counselor, rather than a coach or participant during volley ball games), To establish a profile of Hans' work capacities through vocational testing and to secure appropriate vocational training and experience to enable return to a worker role.
Golden	Woodworking means a lot to Hans. He enjoys working with wood to build furniture and it is a goal he wants to achieve once again. He has the desire to regain the role of woodworker for a productivity as well. With modifications and techniques it can be achieved. Hans values this role and even after going to vocational testing he did not want to be an accountant. Woodworking goals would allow us to develop self efficacy in Hans as well as giving him a means for productivity to be independent once again. This will increase his self confidence and give back a habit.

Table 3: Example of an automatically generated essay and an original student-written essay

the machine-readable format. To reduce the human intervention involved in producing a large amount of training data, we could automate this process by using the ROUGE (Lin, 2004) toolkit. *ROUGE* stands for "Recall-Oriented Understudy for Gisting Evaluation". It is a collection of measures that count the number of overlapping units such as n-gram, word-sequences, and word-pairs between the system-generated summary to be evaluated and the ideal summaries created by humans. We can apply ROUGE to automatically generate extract-based essays given course materials and a set of golden (written by expert human) essays. We can 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 in the document. We can choose the top N 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 can generate essays up to a predefined word limit considering different levels of expertise of the students. In our preliminary experiments, we have generated 214 essays from the given course materials. We have used 20 golden essays[5] in this experiment. The automatically generated essays appeared to be similar in content to that of the original student-written essays. We show an example in Table 3.

Conclusion and Future Work

We used LSA to automatically grade student-written essays. We experimented with different local and global weighting functions applied to the word-by-context matrix. Our experiments revealed that the performance of the LSA model varies with the use of different weighting functions. We also discussed our solutions to reduce human intervention by automating the evaluation framework and the data generation process. In future, we plan to perform large-scale experiments on some other datasets with longer essays and examine how the LSA model's performance varies with respect to different weighting methods.

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 gratefully acknowledge the assistance provided by Dr. Colin Layfield and Dr. Laurence Meadows.

[5]We treated the essays that got the full score of 6 as the golden essays.

References

Briscoe, T., Medlock, B., and Andersen, O. (2010). Automated Assessment of ESOL Free Text Examinations. In *Technical Report UCAM-CL-TR-790 ISSN 1476-2986*, University of Cambridge.

Cederberg, S. and Widdows, D. (2003). Using lsa and noun coordination information to improve the precision and recall of automatic hyponymy extraction. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, pages 111–118. ACL.

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

Clodfelder, K. A. (2003). An lsa implementation against parallel texts in french and english. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts: data driven machine translation and beyond - Volume 3*, pages 111–114. ACL.

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

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

Hacioglu, K., Pradhan, S., Ward, W., Martin, J. H., and Jurafsky, D. (2003). Shallow Semantic Parsing Using Support Vector Machines. In *Technical Report TR-CSLR-2003-03*, University of Colorado.

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

Kakkonen, T., Myller, N., and Sutinen, E. (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)*.

Kakkonen, T. and Sutinen, E. (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.

Kanejiya, D., Kumar, A., and Prasad, S. (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.

Kingsbury, P. and Palmer, M. (2002). From Treebank to PropBank. In *Proceedings of the International Conference on Language Resources and Evaluation*, Las Palmas, Spain.

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

Lin, C. (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.

Moschitti, A. and Basili, R. (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.

Moschitti, A., Quarteroni, S., Basili, R., and Manandhar, S. (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.

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

Pino, J. and Eskenazi, M. (2009). An application of latent semantic analysis to word sense discrimination for words with related and unrelated meanings. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 43–46. ACL.

Powers, D. E., Burstein, J., Chodorow, M., Fowles, M. E., and Kukich, K. (2000). Comparing the validity of automated and human essay scoring. *(GRE No. 98-08a, ETS RR-00-10). Princeton, NJ: Educational Testing Service*.

Rudner, L. M. (1992). Reducing Errors Due to the Use of Judges. *Practical Assessment, Research & Evaluation*, 3(3).

Zhang, A. and Lee, W. (2003). Question Classification Using Support Vector Machines. In *Proceedings of the Special Interest Group on Information Retrieval*, pages 26–32, Toronto, Canada. ACM.