Unsupervised Word Sense Disambiguation for Automatic Essay Scoring

Prema Nedungadi and Harsha Raj

Amrita Vishwa Vidyapeetham
Amritanagar, Tamil Nadu - 641112, India
prema@amrita.edu

Abstract. The reliability of automated essay scoring (AES) has been the subject of debate among educators. Most systems treat essays as a bag of words and evaluate them based on LSA, LDA or other means. Many also incorporate syntactic information about essays such as the number of spelling mistakes, number of words and so on. Towards this goal, a challenging problem is to correctly understand the semantics of the essay to be evaluated so as to differentiate the intended meaning of terms used in the context of a sentence. We incorporate an unsupervised word sense disambiguation (WSD) algorithm which measures similarity between sentences as a preprocessing step to our existing AES system. We evaluate the enhanced AES model with the Kaggle AES dataset of 1400 pre-scored text answers that were manually scored by two human raters. Based on kappa scores, while both models had weighted kappa scores comparable to the human raters, the model with the WSD outperformed the model without the WSD.

Keywords: Latent Semantic Analysis (LSA), SVD, AES, Word Sense Disambiguation.

1 Introduction

With large classrooms, teachers often find it difficult to provide timely and high quality feedback to students with text answers. With the advent of MOOC, providing consistent evaluations and reporting results is an even greater challenge. Automated essay scoring (AES) systems have been shown to be consistent with human scorers and have the potential to provide consistent evaluations and immediate feedback to students [4].

A teacher evaluates student essays based on students' understanding of the topic, the writing style, grammatical and other syntactic errors. The scoring models may vary based on the question, for example, in a science answer the concepts may carry more weight and the grammatical errors may be less important while in a language essay grammar, spelling and syntactical errors may be as important as the content. Hence, for each essay, AES learns the concepts from learning materials and the teachers scoring model from previously scored essays. Most AES systems today that use LSA consider the important terms as

M.K. Kundu et al. (eds.), *Advanced Computing, Networking and Informatics - Volume 1*, Smart Innovation, Systems and Technologies 27,
DOI: 10.1007/978-3-319-07353-8_51,

a bag of words and cannot differentiate the meaning of the terms in the context of the sentence. In order to improve the accuracy of AES, we incorporate unsupervised word sense disambiguation as part of the pre-processing. In contrast with conventional WSD approaches, we did not just take the senses of the target word for scoring; but measure the similarity between gloss vectors of the target word, and a context vector comprising the remaining words in the text fragment containing the words to be disambiguated resulting in better WSD [3].

The rest of the paper is organized as follows: we first present existing approaches to AES. Then we discuss the system architecture of proposed AES system that incorporates WSD. Next we train and test the system and compare the grading accuracy of the AES system that incorporates WSD with the base AES system. Finally, we show using weighted kappa scores that the proposed model has a higher inter-rater agreement than the previous model.

2 Existing Systems

Automated essay scoring systems are very important research area in educational system and may use NLP, Machine Learning and Statistical Methods to evaluate text answers. In this section, we discuss existing essay scoring systems and word sense disambiguation methods.

2.1 Project Essay Grader (PEG)

Project Essay Grade (PEG) is the first research based on scoring essays by computer. It uses measures to evaluate the intrinsic quality of the essay for grading. Proxes denote the estimation of the intrinsic variables such as fluency, diction, grammar, punctuation, etc., Apart from the content, PEG grading is done based on the writing quality.Using multiple regression technique, training in PEG needs to be done for each essay set used. Page's latest experiments achieved results reaching a multiple regression correlation as high as 0.87 with human graders.

2.2 E-Rater

The basic method of E-Rater is similar to PEG. In addition, E-rater measures semantic content by using a vector-space model. Document vector for the essay to be graded are constructed and its cosine similarity is computed with all the pre-graded essay vectors. The essay takes the score of the essay that it closely matches. E-rater cannot detect humour, spelling errors or grammar. It evaluates the content by comparing the essays under same score.

2.3 Intelligent Essay Assessor (IEA)

IEA is an essay grading technique, where a matrix is built from the essay documents, and dimension reduced by the SVD technique. Column vectors are created for each ungraded essay, with cell values based on the terms (rows) from the original matrix. The average similarity scores from a predetermined number of sources that are most similar to this is used to score the essay.

2.4 Corpus-Based Word Sense Disambiguation (WSD)

A corpus-based WSD method uses supervised learning techniques to generate a classifier from training data that are labelled with the sense of the term. The classifier is then used to predict the sense of the target word in novel sentences.

2.5 Graph-Based Word Sense Disambiguation Using Measures of Word Semantic Similarity

This word sense disambiguation in this work is an unsupervised method that uses weighted graph representation for word sense dependencies in text [2]. It uses multiple semantic similarity measures for centrality based algorithm on the weighted graph to address the problem of word sense disambiguation.

3 Automated Essay Scoring with WSD

In our previous work, we had described an automatic text evaluation and scoring tool A-TEST that checks for surface features such as spelling errors and word count and also uses LSA to find the latent meaning of text [5]. In this paper, we discuss the enhancements to the existing system that incorporates word sense disambiguation. Though LSA systems need a large number of sample documents and have no explanatory power, they work well with AES systems [1].

Our AES system is designed to learn and grade essays automatically and first learns important terms from the golden essays or the course materials. Next it uses a set of pre-scored essays that have been given a grade by human raters manually as the training set to create the scoring model. These essays are pre-processed as a list of words or terms with stop-word removal, stemming, lemmatizing, and tokenized.

Algorithm 1. Learn course material with WSD

```
Input: Golden Essays or Course Material
Output: The Reduced Matrix with the correct sense of the terms.

Step1: Preprocess the training essay set (spelling correction, stop word, lemmatizing)
Step2: Extract the sense of each word in every sentence of the essay (WSD).
Step3: Generate the Term-by-document matrix using the output from WSD
Step4: Decompose A into U, V and  (singular value decomposition)
```

3.1 WSD Process

Next we enhance the AES to include a modified version of a WSD algorithm [3] that identifies the sense or meaning of a word in the context of the sentence by considering the sense of every other word in the sentence instead of '?'.The most likely sense is extracted using the word senses from WordNet, and selecting the sense which has highest similarity with the remaining words in the sentence. Once the sense is determined, this sense is fixed and used for determining the sense of the remaining words thus reducing the time complexity of the WSD.

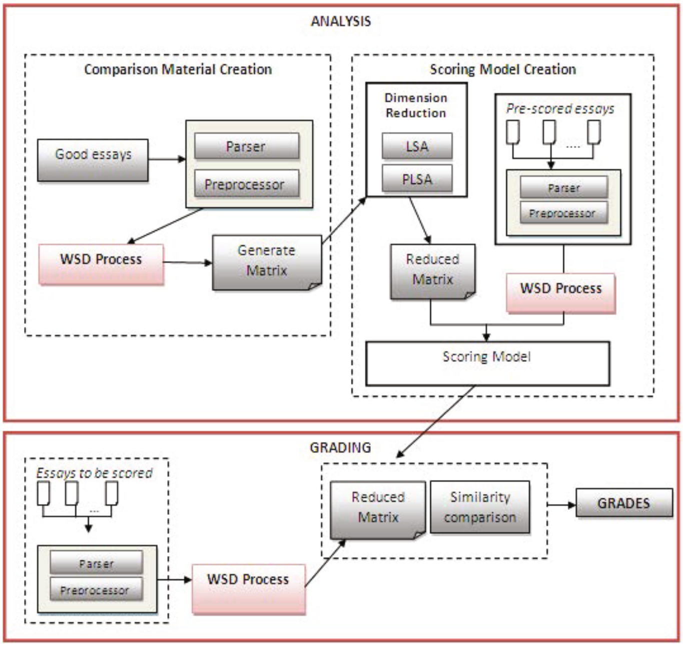

Fig. 1. WSD enhanced AES

The output of the WSD process (Refer Fig. 1) is used by the next phase, LSA to determine the correct of the word used in each sentence of essays [1]. A sense-by-document matrix is created which represents the golden essays, followed by LSA for dimensionality reduction.The document vectors of the pre-scored training essays are created after the pre-processing and WSD process and then compared with the reduced matrix to find the similarity score of the best match. A scoring model is derived using the similarity score, the spelling error and the word count of the essay.

3.2 Scoring Model Using Multiple Regression Analysis

The scoring model was determined using system R, with the similarity score, grade corresponding to the best match, the number of spelling errors and word count.

Scoring Model with WSD

$$Score = diff.in\_spell\_errors * 0.053306 + (Spell - error)^1/4 * 1.008228 + Word\_count * 0.002787 + Similarity\_Measure * 0.277596$$

Scoring Model without WSD

$$Score = diff.in\_spell\_errors * 0.043408 + (Spell - error)^1/4 * 0.793543 + Word\_count * 0.004804 + Similarity\_Measure * 0.174913$$

The last step is the grading phase. Document vectors from the essay to be graded go through the same preprocessing and WSD process and derive the similarity measure of the best match. The scoring model is used to determine the grades for the new essays.

4 Results and Performance Evaluation

We used the Kaggle dataset to test the new AES with WSD and compared the results to the AES without the WSD processing. We tested our model using 304 essays from dataset of 1400 pre-scored essays while the remaining were used to train the AES. There were two human rater scores for each essay that ranged from 1 to 6, where 6 was the highest grade.

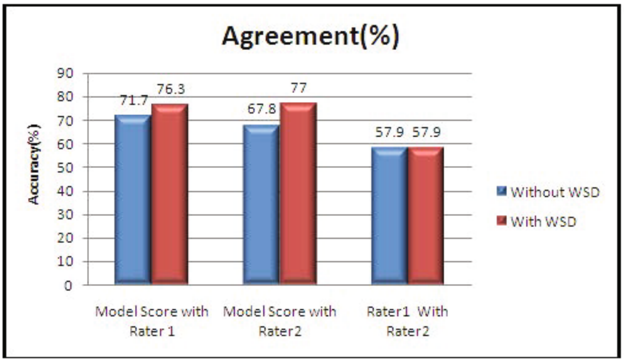

Fig. 2. Agreement between model scores with Rater1 and Rater2

The scores from the first human rater were used to learn the scoring model. The agreement between the model-scores with human raters is shown in Fig. 2. AES with WSD could correctly classify 232 essays from the 304 essays in the testing phase while the AES without WSD could classify 218 out of the 304 essays. The percentage of this is illustrated in the figure. It is interesting to note that the human raters only agreed 57.9% of the time.

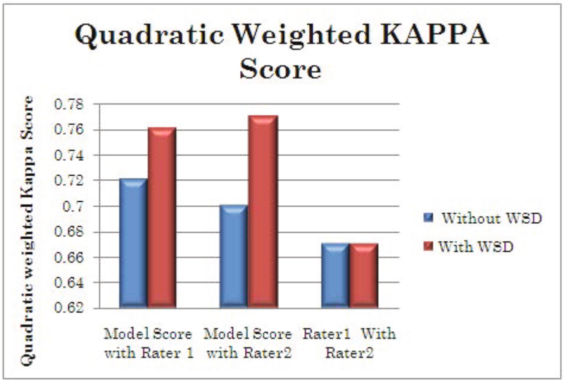

Fig. 3. Quadratic weighted Kappa score

4.1 Kappa Score

While both the models showed good inter-rater agreement, the quadratic weighted Kappa's score show a significant improvement in inter-rater reliability using the AES system with WSD to the base AES system (Refer Fig. 3).

5 Conclusion

This project has presented a new approach for automated essay scoring by taking similarity-based word sense disambiguation method based on the use of context vectors. In contrast with conventional approaches, this did not take the senses of the target word for scoring; the proposed method measures the similarity between gloss vectors of the target word, and a context vector comprising the remaining words in the text fragment containing the words to be disambiguated. This has been motivated by the belief that human beings disambiguate words based on the whole context that contains the target words, usually under a coherent set of meanings. Our results have shown that incorporating WSD to the AES system improves accuracy against existing methods, as evaluated by taking the KAPPA score.

We show that by taking the sense of the word in context to all the other words in the sentence, we improved the inter-rater agreement of our model with the human raters. Though our prediction model worked as well as the manually evaluated teacher model, additional enhancements such as n-grams, grammar specific errors, and other dimensionality methods such as PLSA can further improve the inter-rater accuracy and are planned as further work.

The proposed system is applicable for essay with raw text. In future the proposed work will be extended towards the grading of essays containing text, tables and mathematical equations.

Acknowledgment. This work derives direction and inspiration from the Chancellor of Amrita University, Sri Mata Amritanandamayi Devi. We thank Dr.M

Ramachandra Kaimal, head of Computer Science Department, Amrita University for his valuable feedback.

References

1. Valenti, S., Neri, F., Cucchiarelli, A.: An overview of current research on automated essay grading. Journal of Information Technology Education 2, 319–330 (2003)
2. Sinha, R., Mihalcea, R.: Unsupervised Graph-based Word Sense Disambiguation Using Measures of Word Semantic Similarity. In: IEEE International Conference on Semantic Computing (2007)
3. Abdalgader, K., Skabar, A.: Unsupervised similarity-based word sense disambiguation using context vectors and sentential word importance. ACM Trans. Speech Lang. Process. 9(1) (2012)
4. Kakkonen, T., Myller, N., Sutinen, E., Timonen, J.: Comparison of Dimension Reduction Methods for Automated Essay Grading. Educational Technology and Society 11(3), 275–288 (2008)
5. Nedungadi, P., Jyothi, L., Raman: Considering Misconceptions in Automatic Essay Scoring with A-TEST - Amrita Test Evaluation & Scoring Tool. In: Fifth International Conference on e-Infrastructure and e-Services for Developing Countries (2013)