Pair-Wise: Automatic Essay Evaluation using Word Mover's Distance

Keywords: Automatic Essay Scoring, Word Mover's Distance, Semantic Analysis

Abstract: Automated essay evaluation (AEE) represents not only as a tool to assess evaluate and score essays, but also helps to save time, effort and money without lowering the quality of goals and objectives of educational assessment. Even if the field has been developing since the 1960s and various algorithms and approaches have been proposed to implement AEE systems, most of the existing solutions give much more focus on syntax, vocabulary and shallow content measurements and only vaguely understand the semantics and context of the essay. To address the issue with semantics and context, we propose pair-wise semantic similarity essay evaluation by using the Word Mover's Distance. This method relies on Neural Word Embedding to measure the similarity between words. To be able to measure the performance of AEE, a qualitative accuracy measure based on pairwise ranking is proposed in this paper. The experimental results show that the AEE approach using Word Mover's distance achieve higher level of accuracy as compared to others baselines.

1 INTRODUCTION

Student assessment plays a major role in the educational process and scoring subjective type of questions is one of the most expensive and time-consuming activity for educational assessments. As a consequence, the interest and the development of automated assessment systems are growing.

Automated Essay Evaluation[1] (AEE) can be seen as a prediction problem, which automatically evaluates and scores essay solutions provided by students by comparing them with the reference solution via computer programs (Miller et al., 2013). For academic institutions, AEE represents not only a tool to assess learning outcomes, but also helps to save time, effort and money without lowering the quality of teacher's feedback on student solutions.

The area has been developing since the 1960s when Page and his colleagues (Page, 1966) introduced the first AEE system. Various kinds of algorithms, methods, and techniques have been proposed to implement AEE solutions, however, most of the existing AEE approaches consider text semantics very vaguely and focus mostly on its syntax.

We can assume that most of the existing AEE approaches give much more focus on syntax, vocabulary and shallow content measurements and only limited concerns for the semantics. This assumption follows from the fact that the details of most of the known systems have not been released publicly. To semantically analyze and evaluate documents in these systems, Latent Semantic Analysis (LSA) (Deerwester et al., 1999), Latent Dirichlet Allocation (LDA) (Blei et al., 2003), Content Vector Analysis (CVA) (Attali, 2011) and Neural Word Embedding (NWE) (Mikolov et al., 2013; Kusner et al., 2015) are mostly used.

NWE (Bengio et al., 2003; Mikolov et al., 2013) is similar to other text semantic similarity analysis methods such that LSA or LDA. The main difference is that LSA and LDA utilize co-occurrences of words while NWE learns to predict context. Moreover, training of semantic vectors is resulted from neural networks. NWE models have increased acceptance in recent years because of their high performance in natural language processing (NLP) tasks (Li et al., 2015).

In this work, the Word Mover's Distance (WMD) is utilized which uses word embedding, vector representations of terms, computed from unlabeled data that represent terms in a semantic space in which proximity of vectors can be interpreted as semantic similarity (Mikolov et al., 2013; Kusner et al., 2015). The proposed method measures a distance between individual words from the reference solution and a student answer. To the best of authors' knowledge, this work is the first effort in utilizing WMD for AEE.

The main goal of the proposed WMD based Pair-Wise AEE approach is not to accurately reproduce the human grader's scores, which are varying in their evaluation but to provide acceptable scores and also immediate and helpful feedback. The proposed AEE approach is compared with approaches using LSA, Wordnet and cosine similarity. Experiments showed the proposed WMD based Pair-Wise AEE approach promising such that, in general, it achieved higher evaluation accuracy than the used baseline AEE ap-

[1] Also called Automated Essay Scoring.

proaches.

The rest of this paper is organized as follows: Section 2 reviews existing AEE approaches. In Section 3, the proposed Word Movers Distance based Pair-Wise AEE approach is introduced. Experiments and results are described in Section 4. Section 5 concludes the paper and discusses prospective plans for future work.

2 RELATED WORK

The research on automatically evaluating and scoring essay question answers is ongoing for more than a decade where Machine Learning (ML) and NLP techniques were used for evaluating essay question answers.

Project Essay Grade (PEG) was the first AEE system developed by Ellis Page and his colleagues (Page, 1966). Earlier versions of this system used 30 computer quantifiable predictive features to approximate the intrinsic features valued by human markers. Most of these features were surface variables such as the number of paragraphs, average sentence length, length of the essay in words, and counts of other textual units. PEG has been reported as being able to provide scores for separate dimensions of writing such as content, organization, style, mechanics (i.e., mechanical accuracy, such as spelling, punctuation and capitalization) and creativity, as well as providing an overall score (Shermis et al., 2002; Caryl, 2004). However, the exact set of textual features underlying each dimension as well as details concerning the derivation of the overall score are not publicly disclosed (Ben-Simon and Bennett, 2007; Shermis et al., 2002).

E-Rater (Attali and Burstein, 2006), the basic technique of which is identical to PEG, uses statistical and NLP techniques. E- Rater utilizes a vector-space model to measure semantic content. It examines the structure of the essay by using transitional phrases, paragraph changes, etc., and examines it's content by comparing it's score to other essays. However, if there is an essay with a new argument that uses an unfamiliar argument style, the E-rater will not notice it.

Intelligent Essay Assessor (IEA), based on LSA, is an essay grading technique developed in the late 1990s that evaluates essays by measuring semantic features (Foltz et al., 1999). IEA is trained on a domain-specific set of essays that have been previously scored by expert human raters. IEA evaluates each ungraded essay basically by comparing through LSA, i.e. how similar the new essay is to those it has been trained on. By using LSA, IEA is able to consider the semantic features by representing each essay as a multidimensional vector.

IntelliMetric (Shermis and Burstein, 2003), uses a blend of Artificial Intelligence (AI), NLP and statistical techniques. IntelliMetric needs to be trained with a set of essays that have been scored before by human expert raters. To analyze essays, the system first internalizes the known scores in a set of training essays. Then, it tests the scoring model against a smaller set of essays with known scores for validation purposes. Finally, once the model scores the essays as desired, it is applied to new essays with unknown scores.

AEE systems that use LSA ignore the order of words or arrangement of sentences in its analysis of the meaning of a text because LSA does not have such a feature. A text document in LSA is simply treated as a "bag of words" – an unordered collection of words. As such, the meaning of a text as derived by LSA is not the same as that which could be understood by human beings from grammatical, syntactic relations, logic, or morphological analysis. The second problem is that LSA does not deal with polysemy. This is because each word is represented in the semantic space as a single point and its meaning is the average of all its different meanings in the corpus (Dumais and Landauer, 2008). In this paper, we used the "skip-gram" model of word2vec (Mikolov et al., 2013) to obtain word embedding that learns to predict the context and to train the semantic vectors that is resulted from neural networks to address the issue of word polysemy (Mikolov et al., 2013; Kusner et al., 2015).

3 THE PAIR-WISE APPROACH

The most common way of computing a similarity between two textual documents is to have the centroids of their word embedding and evaluate an inner product between these two centroids (Mikolov et al., 2013; Kusner et al., 2015). However, taking simple centroids of two documents is not a good approximation for calculating a distance between these two documents (Kusner et al., 2015). In this paper, the similarity between individual words in a pair of documents, i.e. the student's answer and the reference (a good) solution, is measured as opposed to the average similarity between the student's answer and the reference solution. Therefore, the Word Mover's Distance (WMD), calculating the minimum cumulative distance that words from a reference solution need to travel to match words from a student answer, was used in this paper.

3.1 Word Mover's Distance

First, it is assumed that text documents are represented by normalized bag-of-words (nBOW) vectors, i.e. if a word w_i appears f_i times in a document, its weight is calculated as

$$d_i = \frac{f_i}{\sum_{j=1}^{n} f_j} \tag{1}$$

where n is the number of unique words in the document. The higher it's weight, the more important the word is. Combined with a measure of word importance, the goal is to incorporate semantic similarity between pairs of individual words into the document distance metric. For this purpose, their Euclidean distance over the word2vec embedding space was used (Kusner et al., 2015; Mikolov et al., 2013). The dissimilarity between word w_i and word w_j can be computed as

$$c(w_i, w_j) = \|x_i - x_j\|^2 \tag{2}$$

where x_i and x_j are the embeddings of the words w_i and w_j, respectively

Let $\mathbf{D}$ and $\mathbf{D}'$ be nBOW representations of two documents D and D', respectively. Let $T \in \mathbb{R}^{n \times n}$ be a flow matrix, where $T_{ij} \geq 0$ denotes how much the word w_i in D has to "travel" to reach the word w_j in D', and n is the number of unique words appearing in D and D'. To transform $\mathbf{D}$ to $\mathbf{D}'$ entirely, we ensure that the complete flow from the word w_i equals d_i and the incoming flow to the word w_j equals d'_j. The WMD is defined as the minimum cumulative cost needed to move all words from D to D', i.e.

$$\min_{T \geq 0} \sum_{i,j=1}^{n} T_{ij}\, c(w_i, w_j) \tag{3}$$

subject to

$$\sum_{j=1}^{n} T_{ij} = d_i, \forall i \in \{1, \ldots, n\}, \sum_{i=1}^{n} T_{ij} = d'_j, \forall j \in \{1, \ldots, n\}$$

The solution is achieved by finding T_{ij} that minimizes the expression in Equation 1. (Kusner et al., 2015) applied this to obtain nearest neighbors for document classification, i.e. k-NN classification which produced outstanding performance among other state-of-the-art approaches. Therefore, WMD is a good choice for semantically evaluating a similarity between documents. The features of WMD can be used to semantically score a pair of texts such that, for example, student's answers and reference solutions.

In this regard, in order to compute the semantic similarity between the student's answer, denoted here by *Sa*, and the reference solution, denoted here by *Rs*, *Sa* is mapped to *Rs* using a word embedding model. Let **Sa** and **Rs** be nBOW representations of *Sa* and *Rs*, respectively. The word embedding model is trained on a set of documents. Since the goal is to measure a similarity between **Sa** and **Rs**, $c(w_i, w_j)$ is redefined as a cosine similarity, i.e.

$$c(w_i, w_j) = \frac{x_i x_j}{\|x_i\| \|x_j\|} \tag{4}$$

Since similarity is used in the Equation 4 instead of distance (Equation 2), Equation 3 is also modified to

$$\max_{T \geq 0} \sum_{i,j=1}^{n} T_{ij}\, c(w_i, w_j) \tag{5}$$

subject to

$$\sum_{j=1}^{n} T_{ij} = d_i, \forall i \in \{1, \ldots, n\}, \sum_{i=1}^{n} T_{ij} = d'_j, \forall j \in \{1, \ldots, n\}$$

3.2 Pair-Wise Architecture

Figure 1 shows the architecture of the proposed WMD based Pair-Wise AEE system.

In preprocessing an essay, the following tasks were performed: tokenization; removing punctuation marks, determiners, and prepositions; transformation to lower-case; stopword removal and word stemming. In the stopword removal step, the words that are in the stop word list (Hípola, 1991) were removed. After removing the stopwords the words have been stemmed to their roots. For stemming the words, M. F. Porter's stemming algorithm (Porter, 1980) was used.

For essay evaluation, the freely available word2vec word embedding which has an embedding for 3 million words/phrases from Google News, trained using the approach in (Mikolov et al., 2013) was used as a word embedding model in the implementation of the WMD based Pair-Wise approach.

4 EXPERIMENT

The experiment was carried out on datasets provided by the Hewlett Foundation at a Kaggle[2] competition for an AEE. There are ten datasets containing student essays from grade ten students. All the datasets were rated by two human raters. The features of the datasets are shown in Table 1.

Five datasets, numbered 1, 2, 5, 9 and 10 in this paper, are provided with the correct, reference solution to which student answers are compared to. In case of the other five datasets (no. 3, 4, 6, 7 and 8)

[2] https://www.kaggle.com/c/asap-sas

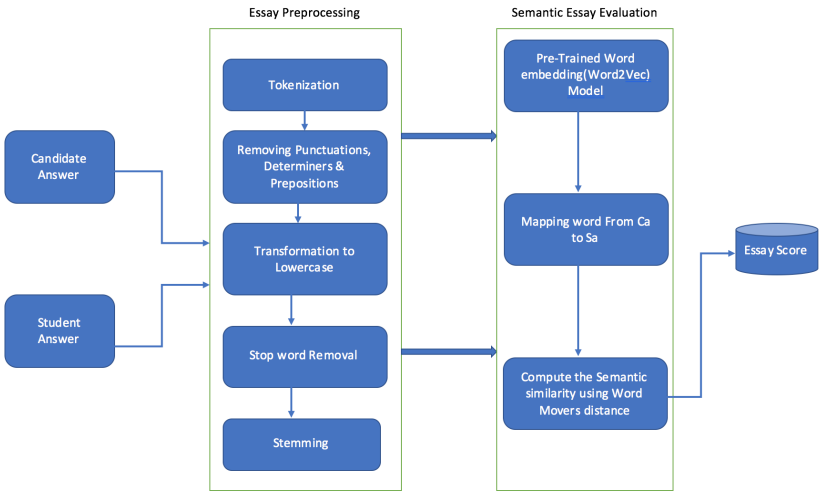

Figure 1: The architecture of the proposed Pair-Wise AEE approach.

Essay Set	Grade Level	Domain	Score range	Average length in words	Training set size	Test set size	Total size
1	10	Science	0-3	50	1672	558	2230
2	10	Science	0-3	50	1278	426	1704
3	10	English, arts	0-2	50	1891	631	2522
4	10	English, arts	0-2	50	1738	580	2318
5	10	Biology	0-3	60	1795	599	2394
6	10	Biology	0-3	50	1797	599	2396
7	10	English	0-2	60	1799	601	2400
8	10	English	0-2	60	1799	601	2400
9	10	Science	0-2	60	1798	600	2398
10	8	Science	0-2	60	1799	599	2398

Table 1: Essay sets used in the experiment and their main characteristics.

the reference solution was created according to the score given by human raters, i.e. ten students' answers which got full score were randomly selected as reference solutions.

Python was used to implement the algorithms discussed. As the Pair-Wise approach is dependent on a word embedding, we used the freely-available Google News word2vec[3] model. Additionally, Scikit-learn[4] and Numpy[5] Python libraries were also used.

The performance of the proposed Pair-Wise approach is compared to that of other three approaches utilizing LSA (Deerwester et al., 1999; Islam and Latiful Hoque, 2010), Wordnet (Atoum and Otoom, 2016; Wan and Angryk, 2007; Zhuge and Hua, 2009) and cosine similarity (Ewees, A et al., 2014; Xia et al., 2015).

[3] https://code.google.com/archive/p/word2vec/
[4] http://scikit-learn.org/
[5] http://www.numpy.org/

4.1 Quantitative Evaluation

The machine score of each essay was compared with the human score to test the reliability of the proposed Pair-Wise approach. Normalized root mean squared error ($nRMSE$) was used to evaluate the agreement between the score given by the Pair-Wise approach as well as baseline AEE algorithms and the actual human scores. The essay scores provided by human raters were normalized to be within $[0,1]$. $nRMSE$ is widely accepted as a reasonable evaluation measure for AEE systems (Williamson, 2009) and is defined as

$$nRMSE(ES) = \left(\frac{\sum_{Sa \in ES} (r(Sa) - h(Sa))^2}{|ES|} \right)^{\frac{1}{2}} \quad (6)$$

where ES is the Essay Set used, $r(Sa)$ and $h(Sa)$ are the predicted rating for Sa by the used AEE approach and the human rating of Sa, respectively. Rating here means how the student answer is similar to the refer-

ence solution. The lower the *nRMSE* the better the performance of the measured approach is.

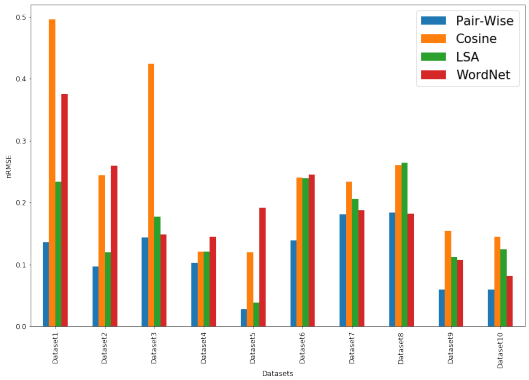

Figure 2: A quantitative comparison using *nRMSE* (Equation 6) of the proposed Pair-Wise AEE and the baselines using LSA, WordNet and Cosine similarity. In case of the datasets no. 3, 4, 6, 7 and 8, the average performance from 10 runs (corresponding to the 10 randomly chosen reference solutions) is reported while in case of the other five datasets (where only the one reference solution indicated in the data is used), the result from one run is reported.

Figure 2 shows the *nRMSE* between the human score and the tested AEE systems for the datasets used in the experiment. Except the Dataset8, where Pair-Wise was performing slightly worse than the winner Wordnet baseline, Pair-Wise was outperforming the baseline approaches.

In case of the datasets 3, 4, 6, 7 and 8, the average values of *nRMSE* from the ten runs corresponding to ten randomly chosen reference solutions are indicated in the Figure 2. To test if the differences between the tested AEE approaches indicated in the Figure 2, in case of these 5 datasets, are statistically significant, the non-parametric Wilcoxon signed-rank test was used. The resulting p-values from these tests are reported in the Table 2 showing that the differences between PairWise and the baselines as well as between the baselines are statistically significant.

4.2 Qualitative Evaluation

nRMSE measures the performance of the tested AEE approaches quantitatively, i.e. by how much the predicted score of an approach differs from the human ratings. Since the proposed and baseline approaches are based on different models, their results might be biased. Thus, a qualitative evaluation measure, named *prank*, referring to "pairwise ranking" is proposed and used in this paper, defined as

$$prank(ES) = \frac{1}{Z} \sum_{Sa_i \neq Sa_j \in ES} \delta(Sa_i, Sa_j) \qquad (7)$$

where $Z = |ES|(|ES| - 1)/2$ is a normalization constant and $\delta(Sa_i, Sa_j) = 1$ if $\big(h(Sa_i) < h(Sa_j)\ \&\ r(Sa_i) < r(Sa_j)\big)$ or $\big(h(Sa_i) > h(Sa_j)\ \&\ r(Sa_i) > r(Sa_j)\big)$ while in cases where $h(Sa_i) = h(Sa_j)$, $\delta(Sa_i, Sa_j) = 1 - |r(Sa_i) - r(Sa_j)|$.

In other words, $\delta(Sa_i, Sa_j)$ results in it's maximal value 1 when the predicted ratings for two student answers Sa_i and Sa_j do not change the human "ranking" of Sa_i and Sa_j w.r.t. their similarities to the reference solution. If the human ranking can not be determined, i.e. the human rated the similarities of Sa_i and Sa_j to the reference solution equally, then the lower the difference between the predicted ratings the better.

As far as the knowledge of the authors goes, none of the state-of-the-art approaches have been evaluated in a qualitative way, only *nRMSE* (or it's variants) was used in all the recent works found.

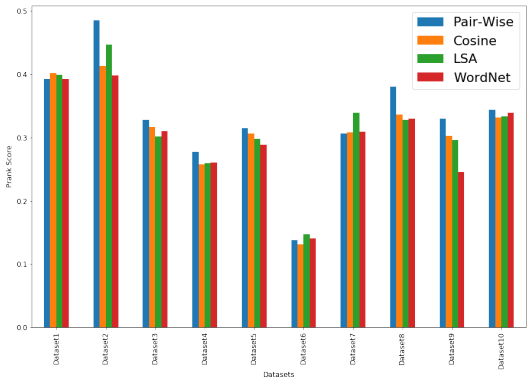

Figure 3: A qualitative comparison using *prank* (Equation 7) of the proposed Pair-Wise AEE and the baselines using LSA, WordNet and Cosine similarity. In case of the datasets no. 3, 4, 6, 7 and 8, the average performance from 10 runs (corresponding to the 10 randomly chosen reference solutions) is reported while in case of the other five datasets (where only the one reference solution indicated in the data is used), the result from one run is reported.

Figure 3 shows the results when measuring the (average) performance of the discussed approaches qualitatively using the proposed *prank* measure. Pair-Wise outperforms the baselines in 7 from the 10 essay sets used for evaluation. In 2 cases, the *prank* score was very close to the winner approaches while only in one case the proposed approach was substantially outperformed by the LSA baseline.

In case of the datasets 3, 4, 6, 7 and 8, the average values of *prank* from the ten runs corresponding to ten randomly chosen reference solutions are indicated in the Figure 3. To test if the differences between the tested AEE approaches indicated in the Figure 3, in case of these 5 datasets, are statistically significant, the non-parametric Wilcoxon signed-rank test was used, as in the case of quantitative evalua-

Datasets	Pair-Wise vs.			LSA vs.		Cosine vs. WordNet
	LSA	WordNet	Cosine	WordNet	Cosine	
Dataset3	0.005	0.005	0.005	0.005	0.005	0.007
Dataset4	0.005	0.005	0.005	0.005	0.005	0.005
Dataset6	0.005	0.005	0.005	0.005	0.005	0.005
Dataset7	0.005	0.005	0.005	0.005	0.005	0.005
Dataset8	0.005	0.005	0.005	0.005	0.005	0.074

Table 2: The p-values resulting from the Wilcoxon signed-rank test between the *nRMSE* results of the proposed AEE and the baselines using LSA, WordNet and Cosine similarity.

Datasets	Pair-Wise vs.			LSA vs.		Cosine vs. WordNet
	LSA	WordNet	Cosine	WordNet	Cosine	
Dataset3	0.005	0.005	0.005	0.878	0.878	0.241
Dataset4	0.012	0.005	0.053	0.012	0.078	0.170
Dataset6	0.006	0.005	0.028	0.005	0.005	0.005
Dataset7	0.016	0.005	0.005	0.053	0.721	0.006
Dataset8	0.005	0.005	0.005	0.332	0.044	0.006

Table 3: The p-values resulting from the Wilcoxon signed-rank test between the *prank* results of the proposed AEE and the baselines using LSA, WordNet and Cosine similarity.

tion, above. The resulting p-values from these tests are reported in the Table 3 showing that the differences between Pair-Wise and the baselines are statistically significant.

5 CONCLUSIONS

In this paper, an automated essay evaluation (AEE) system has been developed using word mover's distance (WMD). During evaluating essays, the system accepts two values. i.e student answer and reference solution. The experimental results showed that there is a significant correlation between the human score and the scores using the proposed Pair-Wise AEE approach. This opens the way for development of AEE systems using semantic features of essays. Such systems can be more helpful for teachers and schools in assessing students using essay type of questions, especially in on-line based learning.

To measure the performance of AEE, a qualitative accuracy measure based on pairwise ranking was also proposed in this work.

The next step of this research is focused to increasing the performance of the proposed system by integrating algorithms that will identify and penalize attempts by students to deliberately fool the system by writing only significant words or phrases in an essay instead of a proper essay and also by creating an own word embedding model. In the future, the proposed approach will be integrated to the web-based AEE system under development by the authors, what will help to test the performance of the proposed Pair-Wise approach in a real-time scenario.

REFERENCES

Atoum, I. and Otoom, A. (2016). Efficient Hybrid Semantic Text Similarity using Wordnet and a Corpus.

Attali, Y. (2011). A Differential Word Use Measure for Content Analysis in Automated Essay Scoring. *ETS Research Report Series*, 36(August).

Attali, Y. and Burstein, J. (2006). Automated Essay Scoring With e-rater® V.2. *The Journal of Technology, Learning, and Assessment*, 4(3).

Ben-Simon, A. and Bennett, R. E. (2007). Toward More Substantively Meaningful Automated Essay Scoring. *The Journal of Technology, Learning, and Assessment*.

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A Neural Probabilistic Language Model. *The Journal of Machine Learning Research*.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Caryl, P. G. (2004). Review of Automated essay scoring: A cross-disciplinary perspective.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1999). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*.

Dumais, T. K. and Landauer, S. (2008). Latent semantic analysis. *Scholarpedia*.

Ewees, A, A., Eisa, M., and Refaat, M. M. (2014). Comparison of cosine similarity and k-NN automated essays scoring. *International Journal of Advanced Research in Computer and Communication Engineering*.

Foltz, P. W., Laham, D., and Landauer, T. K. (1999). Automated Essay Scoring : Applications to Educational Technology. In *World Conference on Educational Multimedia, Hypermedia and Telecommunications (ED-MEDIA)*.

Hípola, P. (1991). G. Salton, Automatic text processing: The Transformation Analysis and Retrieval of Information by Computer. *Procesamiento de Lenguaje Natural*.

Islam, M. and Latiful Hoque, A. S. M. (2010). Automated essay scoring using generalized latent semantic analysis. In *International Conference on Computer and Information Technology*.

Kusner, M. J., Sun, Y., Kolkin, N. I., and Weinberger, K. Q. (2015). From Word Embeddings To Document Distances. *International Conference on Machine Learning*, 37:957–966.

Li, Y., Xu, L., Tian, F., Jiang, L., Zhong, X., and Chen, E. (2015). Word embedding revisited: A new representation learning and explicit matrix factorization perspective. In *IJCAI International Joint Conference on Artificial Intelligence*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositional. In *NIPS*.

Miller, M. D., Linn, R. L., and Gronlund, N. E. (2013). *Measurement and Assessment in Teaching*. Pearson, 11 edition.

Page, E. B. (1966). Grading Essays by Computer: Progress Report. In *Invitational Conference on Testing Problems*.

Porter, M. (1980). The Porter Stemming Algorithm.

Shermis, M. D. and Burstein, J. (2003). Automated essay scoring a cross-disciplinary perspective. *British Journal of Mathematical & Statistical Psychology*.

Shermis, M. D., Koch, C. M., Page, E. B., Keith, T. Z., and Harrington, S. (2002). Trait Ratings for Automated Essay Grading. *Educational and Psychological Measurement*.

Wan, S. and Angryk, R. A. (2007). Measuring semantic similarity using WordNet-based context vectors. In *IEEE International Conference on Systems, Man and Cybernetics*.

Williamson, D. (2009). A framework for Implementing Automated Scoring. *Aera*.

Xia, P., Zhang, L., and Li, F. (2015). Learning similarity with cosine similarity ensemble. *Information Sciences*.

Zhuge, W. and Hua, J. (2009). WordNet-based way to identify Chinglish in automated essay scoring systems. In *International Symposium on Knowledge Acquisition and Modeling*.