Nau et al., 2017

Volume 3 Issue 2, pp. 437 - 450

Date of Publication: 08th September, 2017

DOI-https://dx.doi.org/10.20319/pijss.2017.32.437450

This paper can be cited as: Nau, J., Filho, A., & Passero, G. (2017). Evaluating Semantic Analysis Methods for Short Answer Grading Using Linear Regression. PEOPLE: International Journal of Social Sciences, 3(2), 437-450.

EVALUATING SEMANTIC ANALYSIS METHODS FOR SHORT ANSWER GRADING USING LINEAR REGRESSION

Jonathan Nau
Department of Artificial Intelligence (NIASI), University Center of Brusque, Brusque, SC, Brazil
jonathan.naau@gmail.com

Aluizio Haendchen Filho
Department of Artificial Intelligence (NIASI), University Center of Brusque, Brusque, SC, Brazil
aluizio.h.filho@gmail.com

Guilherme Passero
Department of Artificial Intelligence (NIASI), University Center of Brusque, Brusque, SC, Brazil
guilherme.passer@gmail.com

Abstract

The assessment of free-text answers may demand significant human effort, especially in scenarios with many students. This paper focuses on the automatic grading of short answer written in Portuguese language using techniques of natural language processing and semantic analysis. A previous study found that a similarity scoring model might be more suitable to a question type than to another. In this study, we combine latent semantic analysis (LSA) and a WordNet path-based similarity method using linear regression to predict scores for 76 short answers to three questions written by high school students. The predicted scores compared well to human scores and the use of combined similarity scores showed an improvement in overall

results in relation to a previous study on the same corpus. The presented approach may be used to support the automatic grading of short answer using supervised machine learning to weight different similarity scoring models.

Keywords

Semantic Analysis, Linear Regression, Automatic Grading, Automatic Short Answer Grading

1. Introduction

The use of written answers in the teaching-learning process helps the evaluation of higher cognitive processes, besides developing textual interpretation and production skills. However, the assessment of free-text answers may demand significant effort from the teacher, making it difficult to apply written answers in scenarios with many students.

An automatic short answer grading software can be a very useful pedagogical tool, allowing students to write an answer, receive immediate feedback and rewrite their answers to improve their performance. Technology may also offer advantages over traditional systems, which students await feedback for days, or even weeks, and are susceptible to teachers' subjectivities in the evaluation. Research has been carried out in the search for solutions to automate the process of correction of short answers in virtual learning environments. One of the challenges of this task is the variations of language, that is, the fact that an idea can be expressed through several words.

In this work, we evaluate techniques of natural language processing and semantic analysis for automatic grading of short answers. For this, we use two unsupervised machine learning approaches for textual similarity analysis: (i) latent semantic analysis (LSA), a corpus-based approach that has presented promising results in grading of short answer (Santos & Favero, 2015; Mohler & Mihalcea, 2009); and (ii) a knowledge-based approach that presents similarity metrics between concepts based on WordNet, proposed in (Mohler & Mihalcea, 2009).

The similarity scores obtained by the semantic analysis methods are combined in a linear regression algorithm to predict scores for the answers, comparing it to human scores. With this approach, we achieved promising results using the corpus-based and knowledge-based methods.

2. Techniques of Semantic Analysis and Linear Regression

In this research, we used two different methods to measure the similarity between texts: the Latent Semantic Analysis (LSA) method, with a model built on Brazilian Wikipedia articles,

and a WordNet method, known as the Shortest Path. In addition, the linear regression approach is used to calculate the grades based on similarity scores obtained from LSA and WordNet.

2.1. LSA Model

The LSA (Latent Semantic Analysis) model was first proposed by Landauer & Dutnais (1997). It consists in a statistical-mathematical technique of knowledge abstraction from a corpus, allowing the verification of similarity among words and sentences through their contextual use. The premise is that words which tend to occur together within the same document have some semantic relation.

Below is a summary of the algorithm's behavior, adapted from (Landauer & Dutnais, 1997):

1. *Term/document matrix construction:* It is a representation matrix of the corpus used, with the lines corresponding to words and columns to documents. Initially, the value of the absolute frequency of each word in a particular text is assigned in entry matrix.
2. *Adjustment function:* The absolute frequency of words is adjusted by a function, taken into account the importance of each word (e.g. log / entropy).
3. *SVD (Singular Value Decomposition) of the matrix:* The decomposition into singular values shows the correlations between words.
4. *Reduction to the semantic space:* In order to eliminate rows and columns with the smallest singular values, the matrix is reduced to a dimension between 300 and 500.

Therefore, a vector for semantic representation of a given set of words can be obtained and compared with other semantic vectors. Comparisons are usually made by calculating the cosine of the angle between the vectors.

2.2. WordNet Model

Introduction of WordNet was possible after a pioneering in work from Princeton University (Fellbaum, 1998). It forms a knowledge base where nouns, verbs, adverbs, and adjectives are organized by a variety of semantic relationships. Concepts are presented as lexical words kept within one or more sets of synonyms (synsets). As a common dictionary, WordNet contains word definitions. However instead organized alphabetically, it is rather organized by concept (Leacock & Chodorow, 1998).

Some examples of semantic relationships used by WordNet are hypernymy/hyponymy (is-a), meronymy (is-part-of), synonymous and antonyms. These relationships are associated with words to form a hierarchical structure, which is a useful tool for computational linguistics and natural language processing (Meng, Huang & Gu, 2013). Figure 1 presents the concept "dog" (synset 02084071-n) and some of its relationships in WordNet.

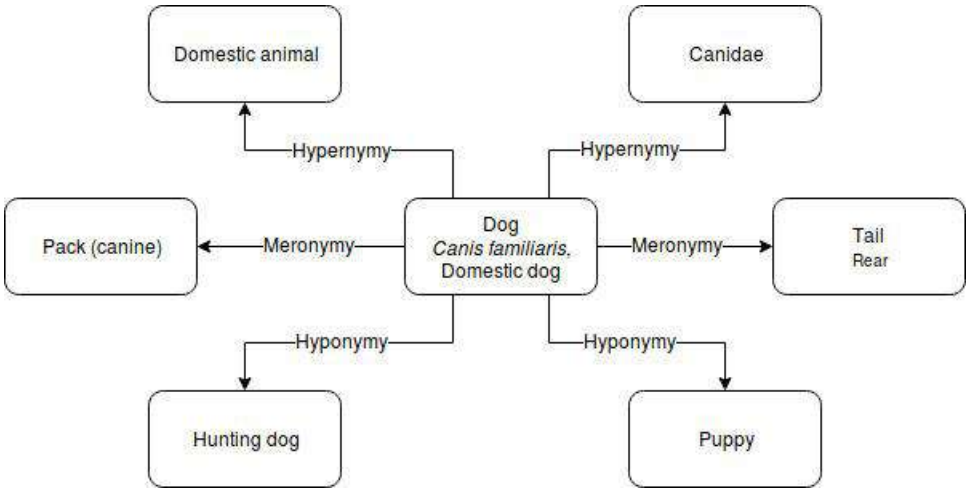

Figure 1: *Concept "dog" and some relationships in WordNet*

Oliveira et al. (2015) compares seven wordnets available for the Portuguese language. According to the author, OWN-PT (OpenWordNet-PT) has free content. It is maintained with machine learning techniques and collaborative human review and has been known for being adopted as the WordNet for Portuguese by the projects FreeLing, Open Multilingual Wordnet and Google Translate. Although there is no precise evaluation method to determine the best WordNet for a context, this study opted for the most popular instance, the OWN-PT, presented in (Paiva, Rademaker & Melo, 2012).

There are several ways of measuring similarity, many of them already proposed in the literature. Passero et al. (2016) describe seven methods to measure semantic similarity between words using WordNet, including path-based and information content-based methods. Based on several tests with the seven WordNet similarity techniques, we chose Shortest Path method, a path-based measure, for its best results when combined with LSA.

The idea of path-based measures is that the similarity between two concepts is a function based on the distance between them and their positioning in WordNet. Shortest Path is the

simplest measure, which considers the shortest distance between two concepts. The formula proposed by (Mohler & Mihalcea, 2009) is:

$$similarity_{ShortestPath} = \frac{1}{Lowest\_distance} \quad (1)$$

2.3. Linear regression

Linear regression is an equation for estimating an expected value (y), by means of values of other variables (x). In many problems, there are two or more related variables, and it may be important to model this relationship. For example, a student's grade may depend on how many hours the student has spent studying for a given grade. Thus, it is possible to construct a model relating the student's grade with a number of study hours.

Figure 1 shows the example of linear regression described. The dots in red indicate the intersection of the hours studied with the grades obtained. The blue line shows the linear regression created according to the data points.

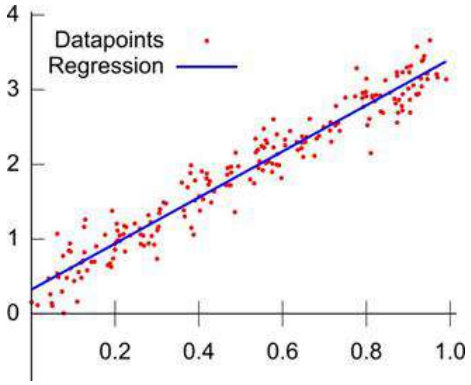

Figure 2: *Example of linear regression*

The general formula of linear regression is:

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_n X_n \quad (2)$$

Where Y is variable explained (dependent), the value one wants to achieve. The alpha (α0) is a constant, which represents the intercept of the line with the vertical axis. The X is the (independent) explanatory variables, which represents the explanatory factors in the equation, weighted by α1 ~ αn. In this study, linear regression was used to combine two similarity scoring methods in a hybrid model to predict short answer scores. Therefore, in our context, Y is the

score to be predicted. The similarity scores represent X, α0 is the bias and α1 ~ αn is the weights learned for each similarity score. For each question, three linear regression models were created.

3. Proposed Approach

The research presented in this paper could be applied to generate knowledge for the solution of specific problems in the linguistic area. The approach comprises five steps: (i) Data collection; (ii) Pre-processing; (iii) Calculation of similarity scores; (iv) Linear regression for grade prediction; and (v) Performance metrics applied to the results.

3.1. Collect and information storage

We used our own application software to collect and store the data. Fourteen high school teachers elaborate questions based on the content seen in the 1st year class. Students from first to 3rd year class answered the questions. The procedures for collecting and storing questions and answers were as follows:

1. Presentation of a seminar for teachers of a secondary school, with a specialist dealing with good practices in the formulation of questions;
2. At the same meeting, teachers were asked to formulate a set of two to four questions, within their respective areas of action, and to enter these questions into the application system developed;
3. Forty-five questions were formulated and then analyzed and validated by the specialist speaker;
4. To compose a test, twenty-one questions from nine areas of knowledge were selected, aiming that the examination time did not exceed two hours;

For the corpus of this research, three discursive questions were selected, two about Portuguese Language and two about Geography. We collected 27, 24 and 25 answers written in the Portuguese language for questions 1, 2 and 3, totaling 76 responses, and the mean number of words per response was 30, 11 and 30, respectively. Two expert teachers gave a score of 0 to 10 for each question. The teachers' reference questions and answers are presented in Table 1.

Table 1: *Questions used in the research corpus*

Geography	**Question 1** **Statement**: Explain the role of DST (daylight saving time) in Brazil.

	Reference answer: Save energy by making better use of light, which extends for longer especially in southern Brazil. This region is more towards the sun due to the axis of inclination of the Earth.
Portuguese **Background** Read the text below: Marcos, 31, was arrested on Sunday afternoon, in the Brusque city, due to an active arrest warrant from the Paraná Court of Justice. The officers of the Tactical Patrol Squad were patrolling the streets when they saw Marcos riding a motorcycle. When consulting the system, they verified the presence of an arrest warrant and that the vehicle was in an irregular situation, which was taken to a detention garage. Marcos was also taken to Advanced Prison Unit (UPA) of Brusque during the afternoon.	. **Question 2** **Statement:** Identify the verbal voice that prevails in the text. **Reference answer:** Passive voice. **Question 2** **Statement:** Justify the use of the predominant verbal voice. **Reference answer:** Passive voice; highlight of the occurring action; generalization of the subject.

In some cases, the teachers' reference answer covered concepts beyond what was asked in the question. In these cases, the responses were adjusted to deal concisely with the problem presented. After validation, reference responses were summarized while maintaining their representativeness. For example, where "The predominant voice is the passive voice", it was reduced to "passive voice".

3.2. Preprocessing

After being collected and stored, the corpus of answers was submitted to preprocessing procedures using CoGrOO 4 (Silva, 2013), with tokenization, named entities recognition, identification of parts of speech (pos-tagging), lemmatization and removal of stop words. The spelling of the answers was manually revised to ensure the correct functioning of CoGrOO.

The complete Wikipedia article base (May / 2016 version) has been translated by an XML parser to a text-only format, keeping the 1.4 million article division. The same debugging procedure applied to the answers was applied to Wikipedia. In this process, the size of the base went from about 5.5GB to 4GB.

3.3. Calculation of similarity scores

The LSA and WordNet methods were used to assess the similarity between the student's answer and the teacher's reference answer.

LSA models were created with the preprocessed Wikipedia base using the open library Semantic Vectors (Widdows & Ferraro, 2008), with the dimensions [200, 250, 300, 350, 400,

450, 500] and the 441,000 most frequent terms (Frequency> = 10). In this study the best results were seen with the 350 dimensions model representing the LSA since it presented.

We use the Apache Jena framework to load OWN-PT data into the main memory (SPARQL Protocol and RDF Query Language). The algorithms used to calculate similarity were adapted from the free WordNet Similarity library (Pedersen, Patwardhan & Michelizzi, 2004).

We calculate the similarity index for the WordNet model considering a vector with the size of the preprocessed reference answer, applying the formula:

$$\boldsymbol{similarity}_{\boldsymbol{answer}} = \frac{\sum \boldsymbol{max}(\boldsymbol{similarity}(\boldsymbol{ref\_word}, \boldsymbol{answer\_word}))}{\boldsymbol{ref\_size}} \quad (3)$$

Each word of the reference response (ref_word) was compared to all words of the preprocessed student answer (answer_word), and the highest similarity found filled a position of the vector. At the end, we obtained the arithmetic average of the values of this vector. This process was applied using the before mentioned shortest path similarity function (similarity). In the absence of similarity, the Levenshtein distance was used. Since the technique considered only nouns and verbs, the adjectives of the reference answer were considered only when found in the student's answer.

3.4. Linear regression for grade prediction

Often, the most difficult part of solving a machine-learning problem may be to find the right estimator for the job. Different estimators are best suited for different types of data and problems. We used the linear regression estimator for its simplicity and high community use. It adjusts a linear model with coefficients to minimize the residual sum of squares between the responses observed in the data set and the answers predicted by the linear approximation.

For meaningful results, it is necessary to validate the entire data set. We then use cross-validation, more specifically the leave-one-out method. This method is used in small samples due to its high level of processing. For a sample size N, a training set for the estimator is created using N-1 examples. The training set is validated in the only example that was left out. The process is repeated N times, each time creating a new training set and disregarding a single example. The error is calculated by the sum of the errors in each test divided by the size of the samples N.

3.5. Measurement metrics of the results.

Three metrics to measure the results were used: (i) Pearson correlation; (ii) mean absolute error; and (iii) mean squared error.

We use Pearson's linear correlation coefficient (r) as a validation metric for the results achieved in the case study. This metric measures the degree of correlation between two sets of values and its result ranges from -1 to 1, with -1 for perfect negative correlation, 0 for no correlation, and 1 for perfect positive correlation between sets. The calculation is done using the following formula:

$$\boldsymbol{r = \frac{covariance(human\ scores, system\ scores)}{stdev(human\ scores) * stdev(system\ scores)}} \quad (4)$$

In addition to linear correlation, we use mean absolute error (MAE) and root mean squared error (RMSE). The main difference between absolute and square error metrics is that the second one penalizes large errors. The formulas applied to MAE and RMSE were:

$$\boldsymbol{mean\_absolute\_error = \frac{\sum abs(human\_score - system\_score)}{number\_of\_scores}} \quad (5)$$

$$\boldsymbol{root\_mean\_squared\_error = sqrt(\frac{\sum(human\_score - system\_score)^2}{number\_of\_scores})} \quad (6)$$

4. Related Works

Several techniques based on corpus and knowledge have been proposed and discussed in the literature. Given a large number of published techniques but with application in specific contexts, new studies are required comparing existing techniques using the same data set (Burrows, Gurevych & Stein, 2015; Ziai, Ott & Meurers, 2012).

In some studies, new parameters are tested in order to refine existing models. In (Santos & Favero, 2015), the standard LSA technique was improved and an accuracy of 84.94% was obtained in a corpus of 349 responses, similar to the 84.93% agreement among the human evaluators.

Mohler & Mihalcea (2009) carried out a similar research, where they explored knowledge and corpus-based techniques on 21 questions and 637 answers written by Computer Science students. The best results were obtained using LSA with a corpus of Wikipedia articles belonging to a specific domain and a refinement based on the best answers (r = 0.5099).

We extend Passero et al. (2016) approach by applying the same corpus-based similarity analysis technique (LSA) and one of the knowledge-based measures they use (Shortest Path) at the same corpus, however under a supervised machine learning model. The use of linear regression is expected to improve the results by deriving a function to translate an LSA and WordNet similarity score into a grade.

Passero et al. (2016) have shown that different similarity models may have better performance in different types of questions. In their study, WordNet-based measures have shown better results in shorter reference answer, on the other hand, LSA performed better in longer answers. Hereafter we present a comparison between the results obtained in Passero et al. (2016) and those obtained in this work.

5. Results and Discussion

We use the Pearson correlation (r) in order to calculate the judge agreement rat, represented by the MAE and RMSE measures. Table 2 presents the values obtained.

Table 2: *Agreement among the evaluators*

Question	Correlation (*r*)	MAE	RMSE
1	0,82	1,792	2,953
2	0,84	1,440	2,300
3	0,71	2,370	3,040
Total	0,70	1,882	2,840

It was also observed that the evaluators provided the same grade with 27 answers (35.53%), differentiated in one or two points in 25 (32.89%), three to five in 21 (27.65%) and six to ten points in 3 (3.95%).

Table 3 presents the results obtained using linear regression for each similarity-scoring model alone and for the two combined.

Table 3: *Summary of the results obtained with the methods used*

	Question									Overall		
	1			2			3					
Model	***r***	***MAE***	***RMSE***	***r***	***MAE***	***RMSE***	***r***	***MAE***	***RMSE***	***r***	***MAE***	***RMSE***
LSA	0,829	1,296	1,599	0,450	3,480	4,015	0,757	1,083	1,581	0,631	1,987	2,646
WordNet	0,713	1,667	2,010	0,910	1,600	1,980	0,804	1,125	1,458	0,840	1,842	1,842

LSA + WordNet	0,811	1,325	1,680	0,902	1,565	2,022	0,839	1,065	1,319	0,867	1,322	1,702

The summary presented in Table 3 indicates that the metrics used have a similar correlation to judge agreement. LSA alone performed better in Question 1, which had a longer explanatory reference answer (32 words) and WordNet had best results in Question 2 and 3.

The two methods combined have shown better results overall. Since linear regression models were created separately for each question, a different weight for LSA and Shortest Path was learned from training sets during cross validation. As expected, LSA similarity scores had a higher weight for Question 1, and lower in questions 2 and 3.

Table 4 compares the results obtained with those presented by Passero et al. (2016). For this, the mean absolute error (MAE) and the root mean squared error (RMSE) in Table 4 were calculated from the author's original data and compared to our model with combined similarity scores. We show the best results achieved in Passero et al. (2016) study based on linear correlation (r), considering all tested similarity models for each and all questions, which were LSA for Question 1, Lin (1998) measure for Question 2, Shortest Path for Question 3 and Wu & Palmer (1994) measure for overall.

Table 4: *Comparison of the obtained results (i) with (Passero et al., 2016) (ii)*

	(i)			**(ii)**		
Question	***r***	***MAE***	**RMSE**	***r***	***MAE***	**RMSE**
1	0,811	1,325	1,680	0,855	1,407	1,700
2	0,902	1,565	2,022	0,872	1,880	2,358
3	0,839	1,065	1,319	0,896	1,167	1,500
All	0,867	1,322	1,702	0,770	2,026	2,580

The results presented in Table 4 show that our hybrid approach performs better than the ones presented by Passero et al. (2016) when considering MAE and RMSE. In regard of linear correlation, our approach had a higher overall score in relation to the previous study but lower scores for Question 1 and 3. A different similarity model performed best for each Question in Passero et al. (2016) and in a real scenario; it may be hard to choose which model could be used. In the other hand, our new approach comprises both LSA and WordNet models and may remain consistent with different types of questions.

6. Conclusion

In this study, we combined existing semantic similarity scoring models to address the task of automatic short answer grading. The presented approach could be used to support the automatic grading of short answers with size ranging from a single sentence to a paragraph. The main limitation of this approach is the need for a training corpus for supervised machine learning.

The semantic similarity models presented in this paper and in the previously mentioned studies are restricted to the analysis of the concepts covered in a text. Thereby, such models would not be regarded in terms of coherence, cohesion, and syntax related properties. For example, "the answer is X" and "the answer is not X" may be treated according to the bag-of-words approach. We then suggest the use of linguistic and syntactic features in future studies to improve the reliability of results.

Our study prescribes a route for several future research projects. An interesting direction is to research techniques for the pragmatic understanding of discourse in order to obtain relevant results for automated essay scoring. We are selecting relevant features to examine the local discourse coherence, using techniques for noun discourse entity identification and coreference resolution. For the prediction of scores, we use linear regression with a nonlinear kernel.

7. Acknowledgement

We would like to thank SBC (Brazilian Computer Society) for authorizing some parts of the article titled "Avaliação do Uso de Métodos Baseados em LSA e WordNet para Correção de Questões Discursivas", published in the XXVII Brazilian Symposium on Computers in Education (SBIE 2016), to be used in this paper. It is referenced as "Passero, G., Haendchen Filho, A. & Dazzi, R. (2016)".

REFERENCES

Burrows, S., Gurevych, I. & Stein, B. (2015). The eras and trends of automatic short answer grading International Journal of Artificial Intelligence in Education. International Journal of Artificial Intelligence in Education, 25(1), 60-117. https://doi.org/10.1007/s40593-014-0026-8

Fellbaum, C. (1998). WordNet: An electronic lexical database. Massachusetts: MIT Press.

Landauer, T. K. & Dutnais, S. T. (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. Psychological Review, 104(2), 211–240. https://doi.org/10.1037/0033-295X.104.2.211

Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet sense similarity for word sense identification. In Fellbaum, C. (Ed.), WordNet: An electronic lexical database (pp. 265-284). Massachusetts: MIT Press.

Lin, D. (1998). An Information-Theoretic Definition of Similarity. Proceedings of International Conference on Machine Learning, 296–304.

Meng, L., Huang, R. & Gu, J. (2013). A Review of Semantic Similarity Measures in WordNet. International Journal of Hybrid Information Technology, 6(1), 1-12.

Mohler, M., & Mihalcea, R. (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: 30 March-3 April 2009, Megaron Athens International Conference Centre, Athens, Greece (pp. 567-575). Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.3115/1609067.1609130

Oliveira et al. (2015). As Wordnets do Português. Oslo Studies in Language, 7(1), 397-424.

Paiva, V., Rademaker, A., & Melo, G. (2012). OpenWordNet-PT: An Open Brazilian Wordnet for Reasoning. In Proceedings of COLING 2012: Demonstration Papers (pp. 353–360). Mumbai, India: The COLING 2012 Organizing Committee.

Passero, G., Filho, A. H., & Dazzi, R. (2016). Avaliação do Uso de Métodos Baseados em LSA e WordNet para Correção de Questões Discursivas. Proceedings of the 17th Brazilian Symposium on Computers in Education, 1136-1145. https://doi.org/10.5753/cbie.sbie.2016.1136

Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004). WordNet::Similarity: Measuring the Relatedness of Concepts. Demonstration Papers at HLT-NAACL. https://doi.org/10.3115/1614025.1614037

Santos, J. C., & Favero, E. L. (2015). Practical use of a latent semantic analysis (LSA) model for automatic evaluation of written answers. Journal of the Brazilian Computer Society, 21(1). https://doi.org/10.1186/s13173-015-0039-7

Silva, W. D. C. M. (2013). Aprimorando o corretor gramatical CoGrOO. Dissertação de Mestrado, Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo.

doi:10.11606/D.45.2013.tde-02052013-135414. Retrieved from: www.teses.usp.br https://doi.org/10.11606/D.45.2013.tde-02052013-135414

Widdows, D., & Ferraro, K. (2008). Semantic Vectors: A Scalable Open Source Package and Online Technology Management Application. Proceedings of the 6th International Conference on Language Resources and Evaluation, 1183-1190.

Wu, Z., & Palmer, M. (1994). Verb semantics and lexical selection. Proceedings of the 32nd annual meeting on Association for Computational Linguistics (pp. 133-138). https://doi.org/10.3115/981732.981751

Ziai, R., Ott, N., & Meurers, D. (2012). Short Answer Assessment: Establishing Links Between Research Strands. In Proceedings of the Seventh Workshop on Building Educational Applications Using NLP (pp. 190-200). Montreal, Canada.