MASTER THESIS

 *Department of Psychology, University of Amsterdam* 

Title: Semantic-Based Algorithm for Scoring Alternative Uses Tests (AUT)

Status (1st draft or revision) : revision **Date:** 20/5/2020

Student(s)

name : Yina Tsai

student ID Card number : 11695986

e-mail address : yina727@gmail.com

Supervisor(s) **Raoul Grasman**

specialization : Psychological Methods (Behavioural Data Science)

second assessor *(if known)* : **Claire Stevenson**

Abstract

Alternative Uses Tests (AUT) is one of the most popular tests for evaluating creativity. Participants were asked to think of unconventional uses for common objects (e.g., brick) within a fixed time. Traditionally, the originality of the responses is scored by several experts, but it is time-consuming and labour-intensive. In this study, we aim to create a semantic-based automatic scoring system for originality scores. We consider different models (ridge regression, LASSO regression, and random forest regressor) and different feature combinations. The results suggest that our algorithm is on par with human expert ratings, and can significantly reduce the time required for expert ratings.

Introduction

Creativity Tasks

Creativity is defined as the ability to create novel and useful responses (Mayer, 1999; Runco & Jaeger, 2012; Said-Metwaly, Noortgate, & Kyndt, 2017). It is considered a valuable asset in solving personnel and organizational problems, and a driver for individual and societal development (Barbot, Besançon, & Lubart, 2015; Said-Metwaly et al., 2017). However, measuring creativity has posed a challenge for researchers (Kaufman, Baer, Cole, & Sexton, 2008; Silvia, 2011). Among the different approaches of measuring creativity, divergent thinking (DT) tests have been most often used (Kaufman, Lee, Baer, & Lee, 2007). The Alternative Uses Tests (AUT) is the most popular DT tests, and it was created by Guilford as part of his Structure of Intellect theory (Guilford, 1956). Participants were asked to think of unconventional uses for common objects (e.g., brick) within a fixed time. The responses are evaluated on four scales: fluency (number of responses), originality (statistical rarity compared to the overall dataset), flexibility (number of different categories), and elaboration (amount of details) (Said-Metwaly et al., 2017).

Despite their popularity, there are some disadvantages of DT tests. Administering the test can be quite long (75 minutes for TTCT), and scoring is also time-consuming. The psychometric properties of DT tests are also debatable. DT tests were considered to have sufficient internal consistency, which ranges from $\alpha = .74$ to $\alpha = .94$ (Baas & Van der Maas, 2015; Cropley, 2000; Silvia, 2011; Silvia et al., 2008). But test-retest reliability is less than sufficient (from $r = .50$ to $r = .93$), and the evidence for validity is often conflicting (Said-

Metwaly et al., 2017). The psychometric problems of DT tests may arise from scoring. For example, originality scores are often confounded with fluency scores, and an answer is more likely to be scored as unique if the sample size is small (Silvia et al., 2008).

The Consensual Assessment Technique (CAT), on the other hand, relies on the expert ratings to avoid the pitfalls of originality scores that are based on the statistical rarity (Amabile, 1982). In CAT, experts are asked to rate the originality of each idea of a respondent on a 5-point Likert scale, and the average score across experts represents the creativity of the respondent. Compared to scoring based on statistical rarity, CAT is more in line with how creativity is assessed in the real world. Thus, the validity of CAT is more desirable because it is evaluated by a panel of experts. However, it may be time-consuming and expensive to assemble a panel of experts for CAT (Kaufman et al., 2007).

An automatic scoring system can reduce the time required for scoring DT tests, and it can also solve the problem of inter-rater reliability. Previous studies suggested that a semantic-based algorithm could yield adequate reliability, although slightly lower than traditional standardized DT scores (Beketayev & Runco, 2016). The semantic-based algorithm in the previous studies was mainly constructed with an associative network such as WordNet (Fellbaum & Christiane, 2005), which is a lexical database that groups English words based on semantic relations. In WordNet, words with similar meaning are grouped into synsets (e.g, “car” and “auto”), and all synsets are connected to other synsets in a hierarchical structure. For example, “motor vehicle” is a hypernymy of “car”, and “sports car” is a hyponymy of “car”.

Word embedding

In this study, we use word embedding to construct the core features. Word embedding is a language modelling technique which takes a text corpus as input and produces a multi-dimensional vector space. Each word in the corpus is represented as a numerical vector, and semantically similar words will have similar representations. With word embedding, the number of features is much smaller than the size of the text corpus (Mandera, Keuleers, & Brysbaert, 2017). In WordNet, the similarity between words is limited to hierarchical representations. But when words are turned into vectors, we can easily add or subtract them. For example, vector(“King”) – vector(“Man”) + vector

("Woman") will result in vector that is similar to the vector representation of "Queen" (Mikolov, Chen, Corrado, & Dean, 2013).

There are two types of models for computing word vectors: CBOW (Continuous Bag of Words) model and skip-gram model. The CBOW model predicts a word based on the neighbouring words in the sentence, and the number of words in the context depends on the window size. On the other hand, the skip-gram model predicts the context based on the input word. For example, given the sentence "Alice was beginning to get very tired of sitting by her sister on the bank and of having nothing to do" and the target word "tired", the CBOW model with the window size of 2 takes the nearby words {get, very, of, sitting}, and uses the sum of their vectors to predict the target. The skip-gram uses "tired" to predict neighbouring words such as "very" and "of" (Figure 1). We use word embeddings as core features because they are more psychologically plausible than count-based models and may be more suitable for behavioural data (Mandera et al., 2017).

Method

Dataset

The data was collected from 729 first-year psychology students at the University of Amsterdam in 2016 and 2017. The Alternative Uses Test (AUT) was used in the study. Participants were asked to generate as many creative uses of "brick" as possible within two minutes.

The originality of each response was scored by two or three expert judges on a 5-point Likert scale: 1 being not at all original, and 5 being very original. Invalid responses received 0 points (e.g., "rectangular", "I'm finished", etc.). Inter-rater reliability was calculated with Pearson correlation, and it ranged from r = 0.53 to r = 0.84 (Table 1). The inter-rater reliability is used for comparability with the semantic-based algorithms.

Table 1. Correlations between expert ratings

	N	Pearson correlation
Rater 1 vs. Rater 2	1553	.53
Rater 1 vs. Rater 3	712	.69
Rater 2 vs. Rater 3	508	.84

The responses and originality ratings were combined into a large table, with each row being a unique rating to a response. The average ratings for each unique response are also calculated (Table 2).

Table 2. Example of originality ratings of "brick"

Response	Originality Rating	Rater	Mean Originality Rating
auto	4	1	3
auto	2	2	3
bakken	3	1	2.67
bakken	2	2	2.67
bakken	3	3	2.67

Feature Construction

All responses are first converted to lowercase, and punctuation and stop words are removed from the sentences. Stop words are the commonly used words (such as "the", "a", "an", "they") that don't provide us with the true meaning of a sentence. A spell checking algorithm was employed to identify and correct misspelt words.

After the responses are cleaned, we first calculated the frequency of a word in the table (term frequency) and its reciprocal (1/frequency) are calculated, which are features commonly used in natural language processing (Mandera et al., 2017).

Each response was mapped to a word vector in 300 dimensions using a fastText pre-trained model for Dutch. This model was trained on Wikipedia and Common Crawl data, using CBOW model with character n-gram of 5 and window size of 5 (Grave, Bojanowski, Gupta, Joulin, & Mikolov, 2018). The Euclidean distance matrix between all pairs of responses was calculated, and two responses with a low semantic distance (e.g., <.2) were combined into one unique response (e.g., huisje bouwen vs. bouw een huis).

In addition to word vectors, we also included two WordNet similarity measures: path similarity and Wu-Palmer similarity (Pedersen, Patwardhan, & Michelizzi, 2004) between "brick" and each response. All responses were translated into English before computing WordNet similarity measures. Path similarity computes the shortest number of edges from one word sense (a discrete meaning of a word) to another, and the words with shorter path distance are more similar. Wu-Palmer similarity takes into account both the depth of the synsets and the depth of the LCS (Least Common Subsumer). Depth is the distance between

word and root, and LCS is the most specific common ancestors of two synsets (Wu & Palmer, 1994), and the Wu-Palmer similarity can be calculated as

$$2\frac{depth(LCS(s1,s2))}{depth(s1)+depth(s2)}$$

For example, the LCS of "kangaroo" (s1) and "bear" (s2) is "mammal" (Figure 1).

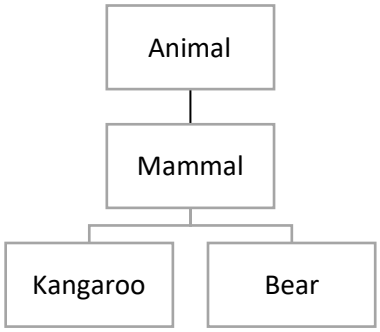

Figure 1. Wu-Palmer similarity considers both the depth of synsets ("kangaroo" and "bear) and the depth of LCS ("mammal").

Overall, we included 304 features in the model: word vector in 300 dimensions, path similarity and Wu-Palmer similarity between "brick" and each response, and the term frequency and its reciprocal (1/frequency). Because the features vary in range and scale, all features are standardized with a mean of 0 and a standard deviation of 1 so that they will contribute equally.

Machine Learning Algorithms

In this study, we consider ridge regression, LASSO regression, and random forest regressor. Because of the large number of features, we can fit ridge and LASSO regression models that constrain some coefficient estimates towards zero. The ridge coefficient estimates $\beta_0, \beta_1, \beta_2, \ldots\ldots, \beta_p$ are the values that minimize

$$\sum_{i=1}^{n}(y_i-\beta_0-\sum_{i=1}^{p}\beta_j x_{ij})^2+\lambda\sum_{i=1}^{p}|\beta_j| \sum_{i=1}^{n}(y_i-\beta_0-\sum_{i=1}^{p}\beta_j x_{ij})^2+\lambda\sum_{i=1}^{p}|\beta_j| = \text{RSS} + \sqrt{\lambda\sum_{i=1}^{p}\beta_j^2}\lambda\sum_{i=1}^{p}\beta_j^2$$

In LASSO regression, the coefficient estimates $\beta_0, \beta_1, \beta_2, \ldots\ldots, \beta_p$ are the values that minimize

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{i=1}^{p}\beta_i x_{ij})^2 + \lambda\sum_{i=1}^{p}|\beta_j| \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{i=1}^{p}\beta_i x_{ij})^2 + \lambda\sum_{i=1}^{p}|\beta_j| = \text{RSS} + \lambda\sum_{i=1}^{p}|\beta_j| \lambda\sum_{i=1}^{p}|\beta_j|$$

Ridge regression includes all features while shrinking coefficient estimates toward zero, while LASSO regression performs feature selection by causing some coefficient estimates to be exactly zero (James, Witten, Hastie, & Tibshirani, 2013).

Another approach is to apply tree-based algorithms. A regression tree is built through recursive binary splitting, which splits data into branches in a top-down approach and each split results in two new branches. Although the regression tree can be easily understood with its graphical display, the predictive accuracy tends to be lower than other regression methods because of overfitting. As an alternative, the random forest can reduce overfitting by building an ensemble of regression trees, and only a subset of features is considered at each split (James et al., 2013).

The dataset is divided into training and test sets in the ratio of 7:3, and we perform cross-validation on the training set to ensure the reliability of the algorithm. We use the features to predict the mean originality ratings, and R^2 is computed to determine which algorithm is more suitable. The interpretability of the algorithm is also taken into consideration when selecting the model.

Random Forest Hyperparameter Tuning

To find the optimal hyperparameters for the random forest regressor, we run cross-validation with different combinations of hyperparameters. But cross-validation on the training set can lead to overfitting. Thus, it is important to use the hyperparameters that optimize model performance but also minimize overfitting. Three hyperparameters that may be most important for random forest regressor:

(1) The maximum depth of the tree: deeper trees capture more information about the data, but it can lead to overfitting when the depth value increases.

(2) The number of estimators: it represents the number of trees in the forest. Increasing the number of estimators can decrease overfitting, but it will eventually reach a plateau. Increasing the number of trees will also increase the processing time.

(3) The maximum number of features: it represents the number of features to consider when splitting. Increasing the maximum number of features will increase the processing time.

Results

Model performance

The performance of ridge regression, LASSO regression and random forest regressor models are comparable to inter-rater reliability (r = 0.53 to r = 0.84) of expert ratings, and random forest regressor performs better than ridge and ridge and LASSO regressions (Table 3).

Table 3. Model performance

Model	**R^2 training set**	**R^2 test set**
Ridge regression	0.63	0.57
LASSO regression	0.63	0.57
Random forest	0.85	0.75

If we plot the mean expert ratings against predictions, we can see that random forest regressor (Figure 4) has better performance than ridge (Figure 2) or LASSO regression (Figure 3), and the performance of ridge and LASSO regressions are similar. When the mean expert rating is lower than 1 or higher than 3, we see some line patterns in the graph. The prediction is inaccurate within that range because of the limited number of data points that were available for training.

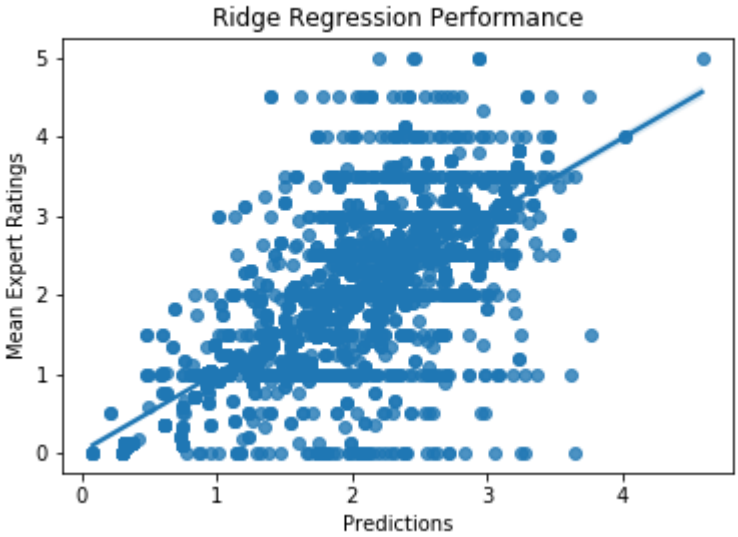

Figure 2. Correlation between ridge regression predictions and mean expert ratings.

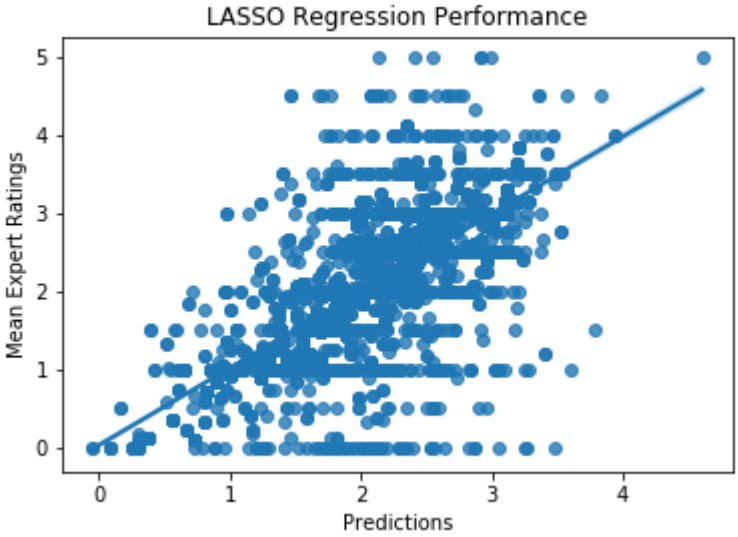

Figure 3. Correlation between LASSO regression predictions and mean expert ratings.

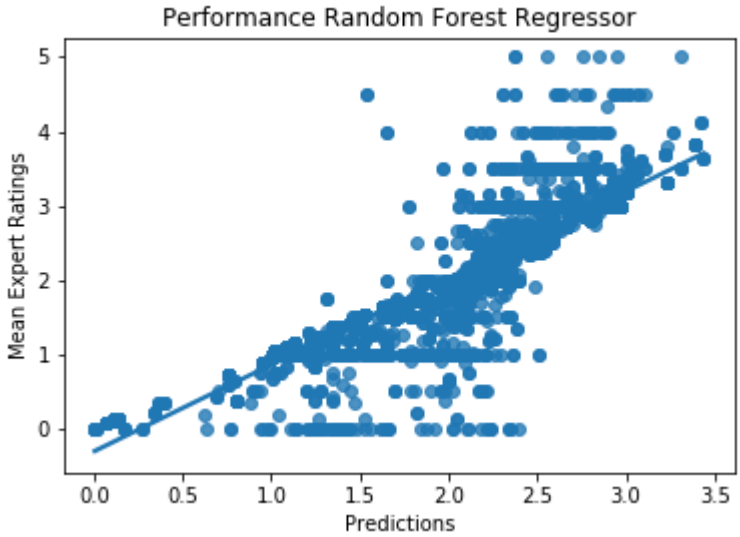

Figure 4. Correlation between random forest regressor predictions and mean expert ratings.

Most of the expert ratings are between 1-3, and there are very few instances with ratings of 5 (Figure 5). This may be the reason that our model does not predict a rating of 5 even when the mean expert rating is 5. For each response, we only had two or three expert ratings. Thus, the mean expert ratings will be more discrete than continuous.

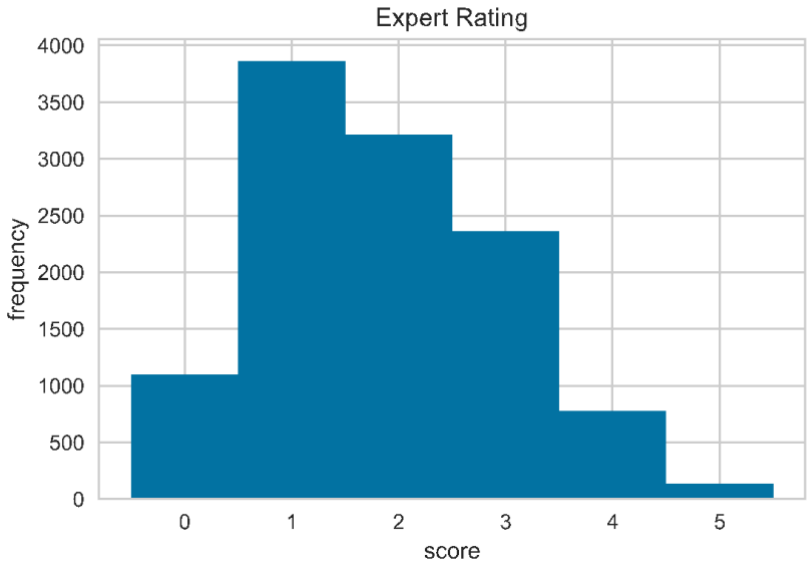

Figure 4. Distribution of human expert ratings.

Feature Importance

There are 304 features in the model, and most features have small coefficient estimates. When comparing feature importance between ridge regression (Figure 5), LASSO

regression (Figure 6), and random forest regressor (Figure 7), the reciprocal of frequency (1/frequency) and frequency have relatively high importance in the models. The semantic vectors in 300 dimensions are also important for the model fit, while some dimensions may contribute more than others. WordNet similarity measures, such as path similarity and Wu-Palmer similarity, are less important for the model fit.

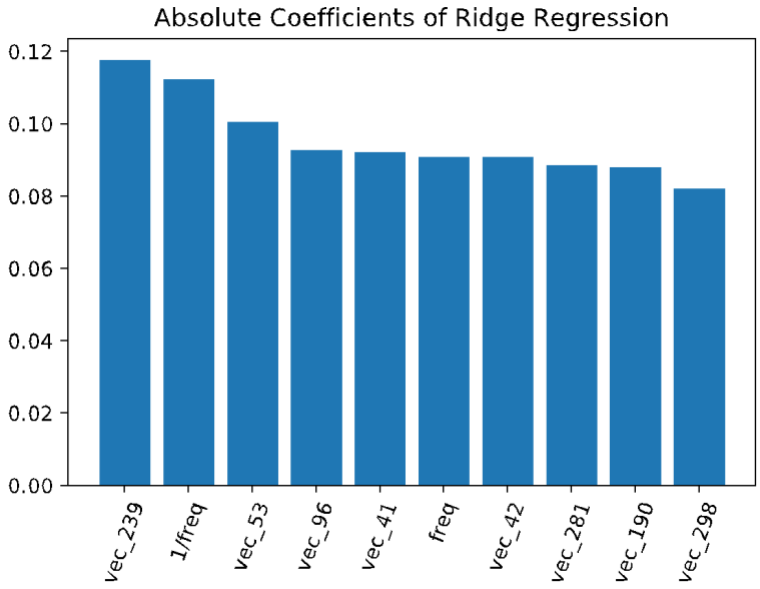

Figure 5. The absolute value of coefficient estimates in ridge regression.

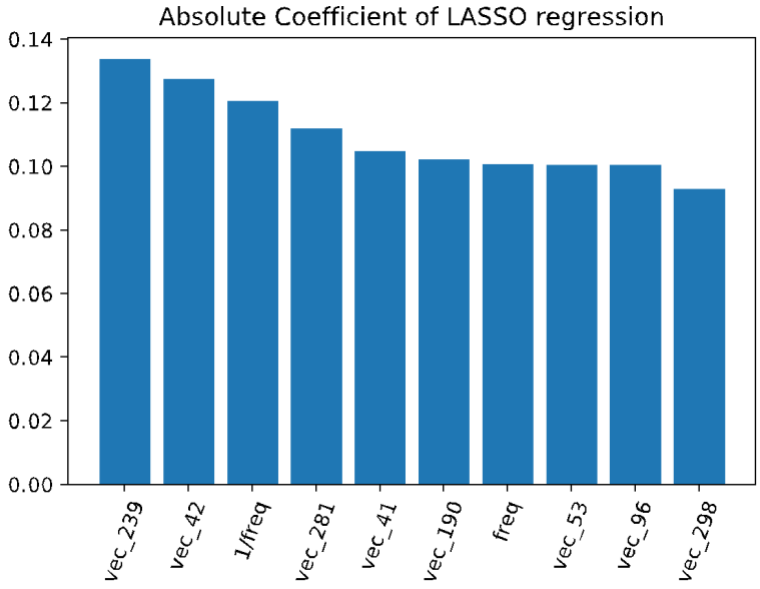

Figure 6. The absolute value of coefficient estimates in LASSO regression

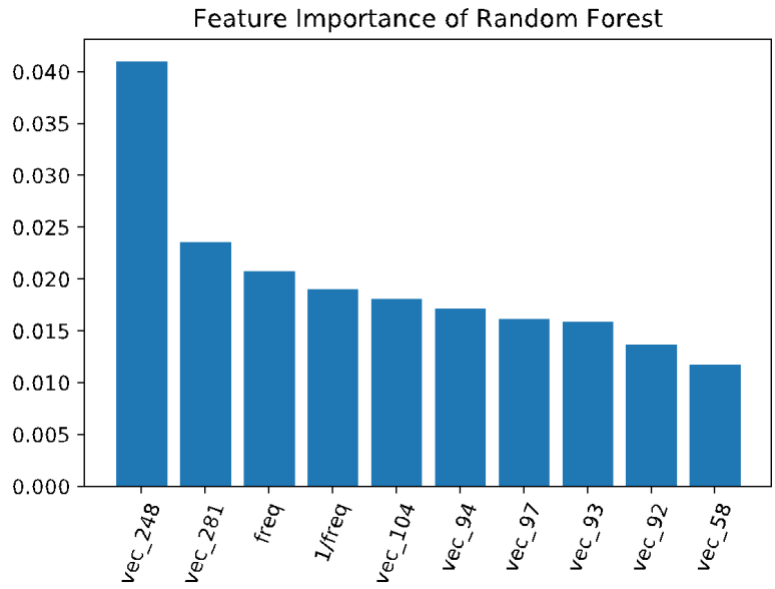

Figure 7. Feature importance in random forest regressor

To better understand how the word vectors and words relate, we plot a random set of responses against the two most important coordinates 248 and 281 (Figure 8). Words that are closer together are more similar in those two vector spaces. For example, “stad” (“city”) and “straat” (“street”) are more similar than “gooien” (“to throw”). However, because we’re only seeing two word vectors out of 300 vectors, the semantic differences between those words cannot be fully explained by just two vectors.

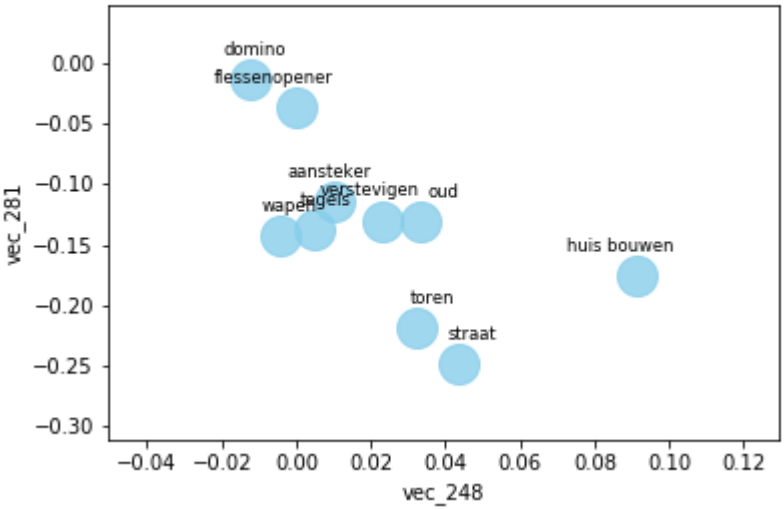

Figure 8. Random words in the word vector space

Discussion

Overall, our semantic-based scoring algorithm is on-par or better than human expert ratings. Among the different algorithms, random forest regressor yields the best performance. However, because of the large amounts of features, random forest regressor requires a longer processing time than ridge or LASSO regression. In terms of model interpretability, random forest is more complex than ridge and LASSO regression and less interpretable. In our study, ridge regression and LASSO regression have similar performance, but LASSO regression provides a simpler model because some coefficient estimates are exactly equal to 0. Therefore, we recommend using random forest regressor if there is sufficient processing power. But if there is a time limitation, LASSO regression can produce satisfactory performance and a more interpretable model.

Based on our results, word vectors can be useful in predicting human expert ratings of originality scores, and some coordinates (e.g., 248, 281) may be more important than other coordinates. Including count-based features such as frequency and its reciprocal also improved our model predictions. This can be explained by the fact that human experts are likely to take statistical rarity into account when rating originality.

In the future, we can also try a combination of models. For example, we can use LASSO regression to remove features with coefficient estimates of 0 and run the random forest regressor based on the reduced features. That way, we can decrease the processing power and time required for the random forest and may be able to improve prediction.

Limitations

One limitation of the current data set is that there are very few instances for high ratings. If we can combine more expert ratings on the same responses, we will have better training data, and may be able to improve our predictions.

Although word embedding may be useful in predicting in predict human expert ratings, it is hard to interpret how each word is represented in a particular word vector space. Thus, it's hard to conclude why some coordinates are more important than the others.

In the study, we only focus on the originality ratings of alternative usage of "brick" because it has the largest number of responses. But there are also other items, such as paperclip or fork, included in the AUT data. It will be interesting to see if our algorithm can predict originality ratings on different items.

In conclusion, our semantic-based algorithm is on-par with human expert ratings. It reduces the time required for expert ratings, and it is less labour-intensive. In the future, it is worth considering using a combined algorithm. We can conduct feature selection with LASSO regression, and use the selected features on random forest regressor.

Reference

Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology*, *43*, 997–1013.

Barbot, B., Besançon, M., & Lubart, T. (2015) Creative potential in educational settings: its nature, measure, and nurture. *Education 3-13,* 43(4), 371-381. https://doi.org/10.1515/ctra-2017-0013

Baas, M. & Van der Maas, H.L.J. (2015). De (on)mogelijkheid van een valide meting van creatief potentieel voor selectiedoeleinden [Selecting creative people: Methodological and practical considerations]. *Gedrag en Organisatie, 28(2),* 78-97. https://doi.org/10.5553/GenO/092150772015028002002

Beketayev, K., & Runco, M. A. (2016). Scoring divergent thinking tests by computer with a semantics-based algorithm. *Europe's Journal of Psychology*. https://doi.org/10.5964/ejop.v12i2.1127

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *TACL, 5*, 135-146.

Cropley, A. J. (2000). Defining and measuring creativity: Are creativity tests worth using? *Roeper Review, 23*(2), 72-79. https://doi.org/10.1080/02783190009554069

Fellbaum, Christiane (2005). WordNet and wordnets. In: Brown, Keith et al. (eds.), Encyclopedia of Language and Linguistics, Second Edition, Oxford: Elsevier, 665-670.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning Word Vectors for 157 Languages. Paper presented at the 11th International Conference on Language Resources and Evaluation. arXiv:1802.06893v2 [cs.CL]

James, G., Witten, D., Hastie, T., & Tibshirani, R (2013). *An Introduction to Statistical Learning: with Application in R.* New York, NY: Springer.

Kaufman, J. C., Lee, J., Baer, J., & Lee, S. (2007). Captions, consistency, creativity, and the Consensual Assessment Technique: New evidence of reliability. *Thinking Skills and Creativity, 2*(2), 96-106. https://doi.org/10.1016/j.tsc.2007.04.002

Kaufman, J. C., Baer, J., Cole, J. C., & Sexton, J. D. (2008). A comparison of expert and nonexpert raters using the Consensual Assessment Technique. *Creativity Research Journal, 20*(2), 171-178. https://doi.org/10.1080/10400410802059929

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language, 92*, 57-78. https://doi.org/10.1016/j.jml.2016.04.001

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *ICLR Workshop*.

Mumford, M. D. (2003). Where have we been, where are we going? Taking stock in creativity research. *Creativity Research Journal, 15*(2-3), 107-120. https://doi.org/10.1080/10400419.2003.9651403

Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004). WordNet::Similarity - Measuring the relatedness of concepts. In *Proceedings of the National Conference on Artificial Intelligence*.

Runco, M.A., & Jaeger, G.J. (2012). The standard definition of creativity. *Creativity Research Journal, 24*(1), 92-96. https://doi.org/10.1080/10400419.2012.650092

Said-Metwaly, S., Noortgate, W., & Kyndt, E. (2017). Approaches to Measuring Creativity: A Systematic Literature Review, *Creativity. Theories – Research - Applications*, *4*(2), 238-275. https://doi.org/10.1515/ctra-2017-0013

Silvia, P.J. (2011). Subjective scoring of divergent thinking: Examining the reliability of unusual uses, instances, and consequences tasks. *Thinking Skills and Creativity, 6*(1), 24-30. https://doi.org/10.1016/j.tsc.2010.06.001

Silvia, P. J., Winterstein, B. P., Willse, J. T., Barona, C. M., Cram, J. T., Hess, K. I., … Richard, C. A. (2008). Assessing creativity with divergent thinking Tasks: exploring the reliability and validity of new subjective scoring methods. *Psychology of Aesthetics, Creativity, and the Arts*, *2*(2), 68-85. https://doi.org/10.1037/1931-3896.2.2.68

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