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

Yina Tsai

June 9th, 2020

Table of Contents

Creativity

- Creativity is defined as the ability to create novel and useful responses (Said-Metwaly, Noortgate, & Kyndt, 2017).
- It is important for problem solving (Barbot, Besançon, & Lubart, 2015).
- But measuring creativity has been a challenge for researchers.
- The Alternative Uses Tests (AUT) is the most popular divergent thinking (DT) test: participants are asked to think of unconventional uses of common objects within a fixed period of time (Guilford, 1956).

Scoring Creativity

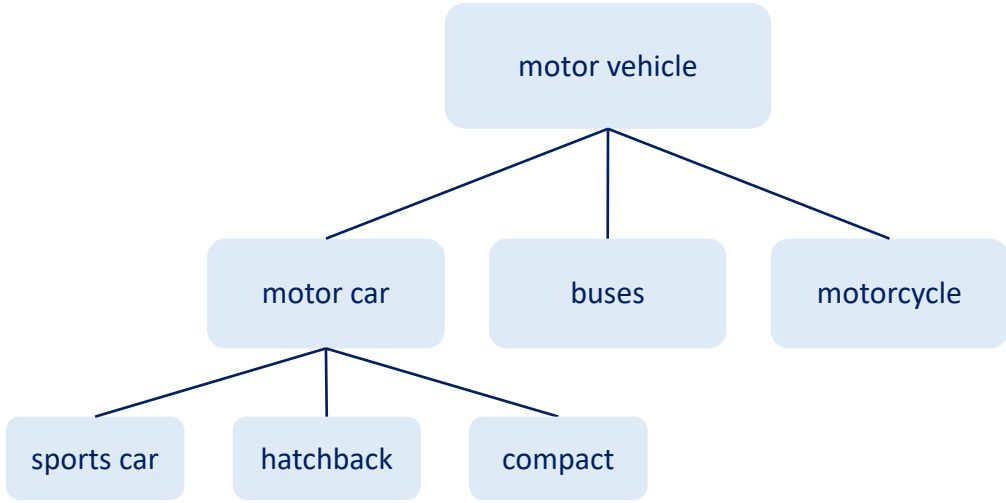
- Evaluation of AUT: fluency, originality, flexibility, elaboration (Said-Metwaly et al., 2017).
- Disadvantages of DT tests: time-consuming, insufficient test-retest reliability, conflicting evidence for validity.
- Consensual Assessment Technique (CAT) relies on expert ratings of originality scores: better validity but still time-consuming and expensive.

Research Questions

- How can we reduce the time for scoring?
- How can we create an automatic scoring system to predict expert ratings?

Automatic Scoring System

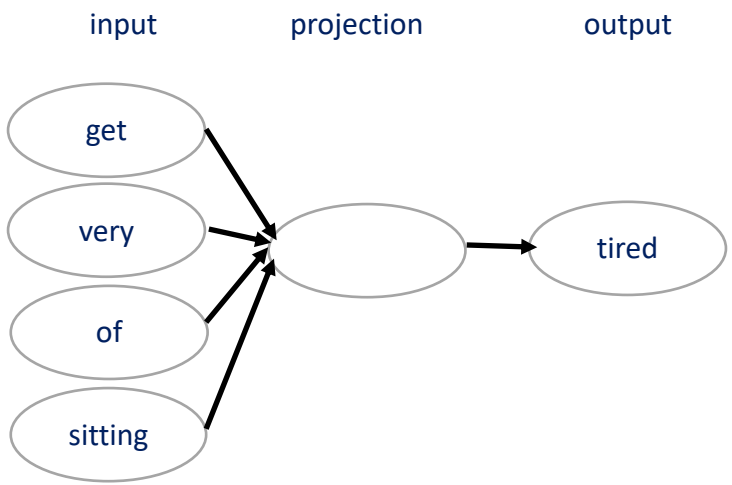
- An automatic scoring system can reduce the time for scoring, and has proven to yield adequate reliability (Beketayev & Runco, 2016).
- The semantic-based algorithm in previous studies was mainly based on associative networks such as WordNet.



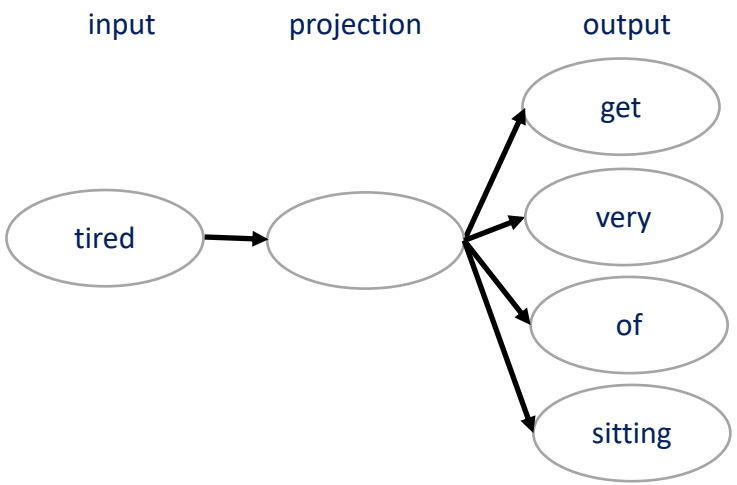
Example of WordNet

Word Embedding

- We chose to use word embedding to construct the core features.
- Each word is represented as a numerical vector in the multi-dimensional space.
- Numerical manipulations on word vectors: king – man = queen - woman
- Two types of models: CBOW and skip-gram.



CBOW model



Skip-gram model

Methods: Dataset

- Participants: 729 psychology students at UvA in 2016 and 2017.
- Task: to think of as many creative uses of “brick” as possible within 2 minutes.
- Originality rating: 2-3 experts on 5-point Likert scale.
- Inter-rater reliability: r = 0.53 to r = 0.84

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

Methods: Feature Construction

Pre-processing:

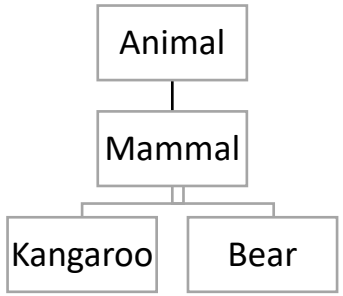
- converted responses to lowercase
- removed stop words and punctuation
- performed spell-check

Word Embedding (fastText): each response was mapped to a word vector of 300 dimensions.
Combine responses with short semantic distance (e.g., huisje bouwen vs. bouw een huis)

Count-based features: frequency, 1/frequency

WordNet features: path similarity, Wu-Palmer similarity

(e.g., $2\frac{depth\ (\text{“}mammal\text{”})}{depth(\text{“}kangeroo)+depth(\text{“}bear\text{”})}$)



Methods: Machine Learning Algorithms

- Use 304 features to predict mean originality ratings.
- Models:
  - Ridge regression: $\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}|\beta_j|$ = RSS + $\lambda \sum_{j=1}^{p}\beta_j{}^2$
  - LASSO regression: $\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}|\beta_j|$ = RSS + $\lambda \sum_{j=1}^{p}|\beta_j|$
  - Random forest: an ensemble of regression trees, and only a subset of features are considered at each split.
- Training : testing = 7 : 3 split
- Cross-validation on training set.

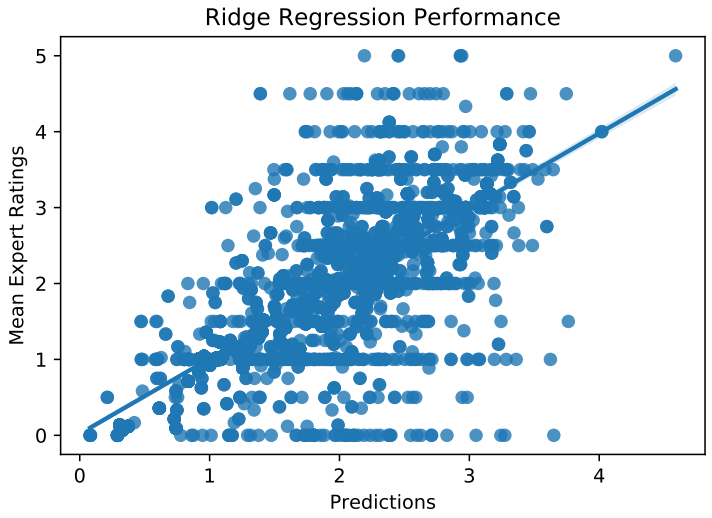
Methods: Hyperparameter Tuning of Random Forest Regressor

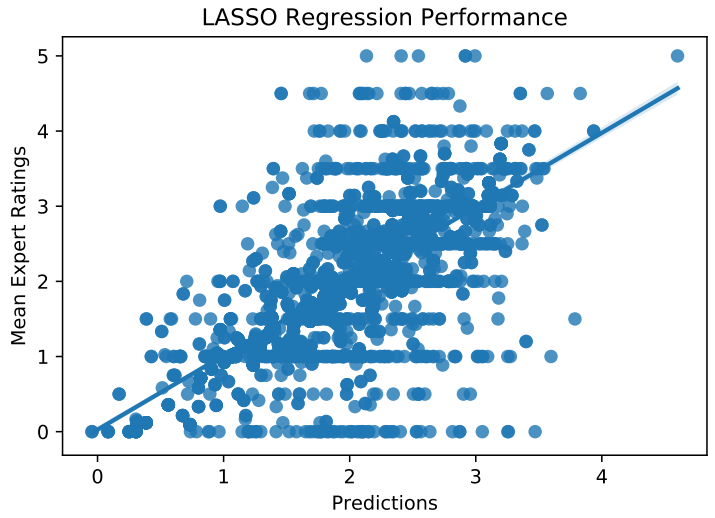
- The maximum depth of the tree: deeper trees capture more information about the data, but it can lead to overfitting.
- The number of estimators: the number of trees in the forest. Increasing the number of trees can reduce overfitting but will reach a plateau, and it also increases processing time.
- The maximum number of features: the number of features to consider when splitting. Increasing it will increase processing time.

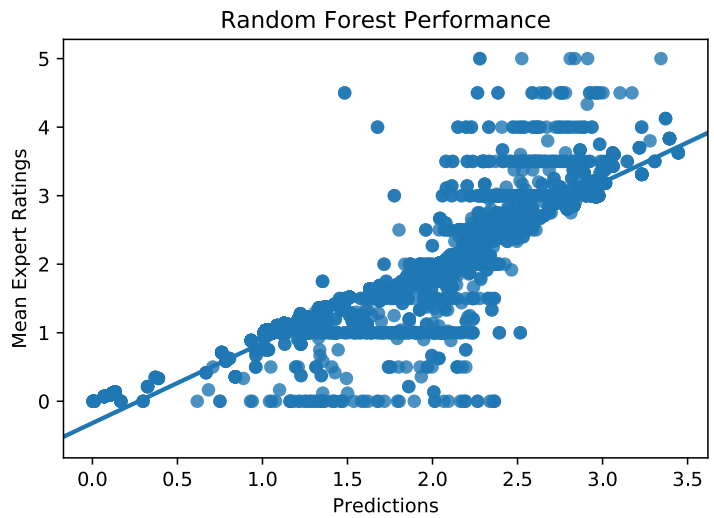
Results: Model Performance

- The performance of the 3 models are comparable to inter-rater reliability of expert ratings (r = 0.53 to r = 0.84).

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

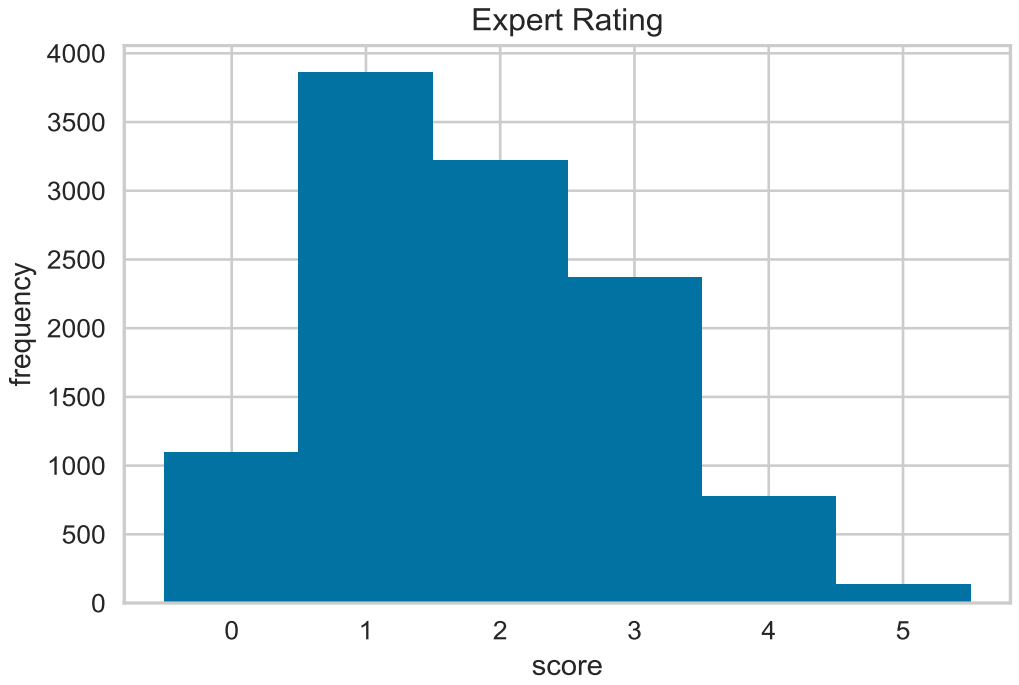






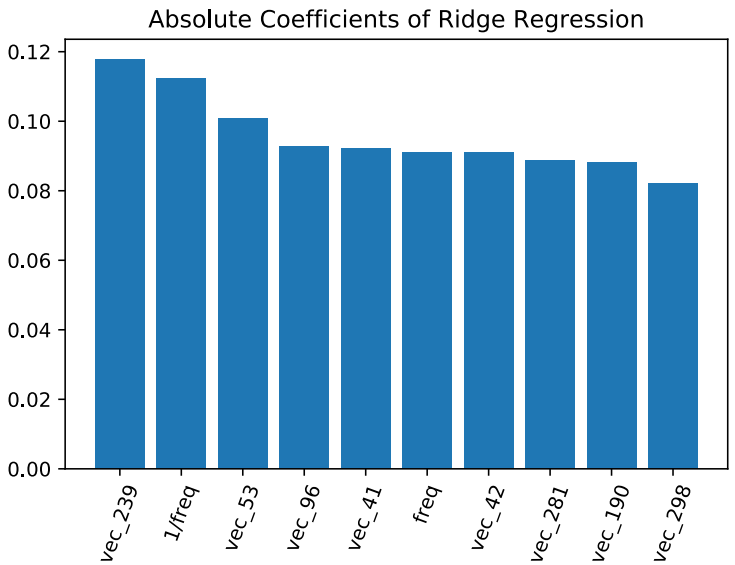
Results: Expert Ratings

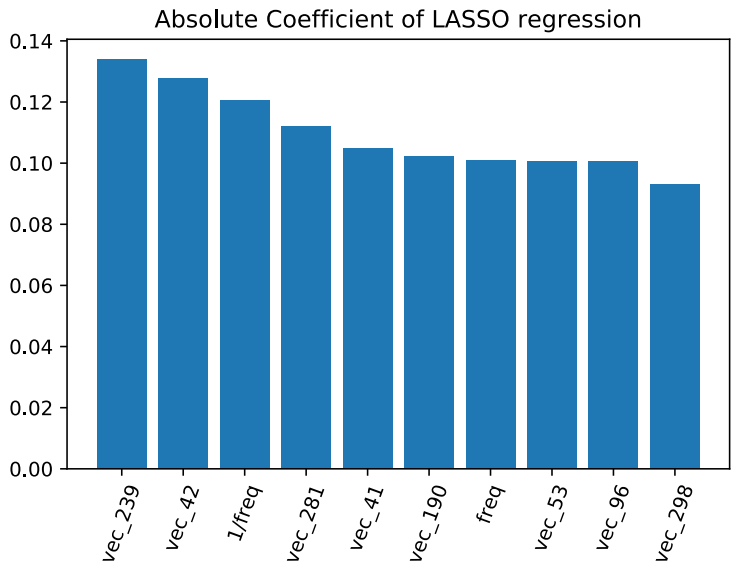
- The models don’t predict 5 when mean expert ratings is 5.
- Most expert ratings are between 1-3, with very few instances of 5.

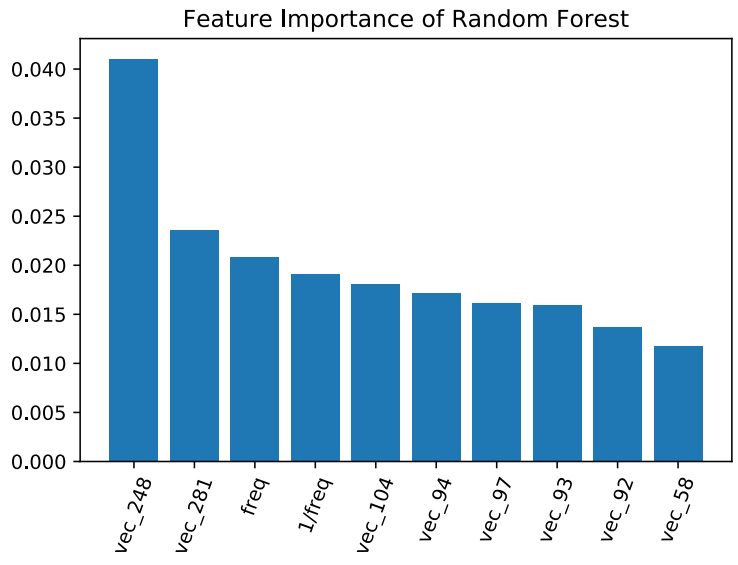


Results: Feature Importance

- Frequency and 1/frequency are important in predicting mean expert ratings.
- Some word coordinates are more important, but it's hard to interpret the coordinates.
- WordNet similarities are not within the 10 most important features.







Discussion

- Our algorithms are on-par or better than inter-rater reliability of human expert ratings.
- Random forest regressor yields better performance but also requires more processing power and time.
- LASSO regression provides a simpler model because some coefficients estimates are equal to 0.
- In the future, we should try combining different models.

Limitations

- Very few instances of higher expert ratings.
- Word embedding may be useful for predicting expert ratings, but it’s hard to interpret each coordinate.
- We only focused on the creative uses of “brick” in this study, but it’ll be interesting to know if the model still holds for other items.

Questions?