Automatic Argument Evaluation

Jason Liang and Eric Wang

Introduction

- Automatic essay scoring (AES) is in widespread use, but very controversial
- Work based on paper “Modeling Argument Strength in Student Essays” by Persing and Ng
- Train model to score essay based on argument strength
- Corpus: 1800 essays from Kaggle written by 10th graders
- Each essay scored by two graders from 1-6 (1 is worst, 6 is best)

Baseline Systems

- Most frequent score

Score	1	2	3	4	5	6
Number of Essays	48	308	1487	1594	148	15

- Features based on discourse connectives
  - Words such as “once”, “since” that describe logical connections between discourse elements

Features in Persing and Ng

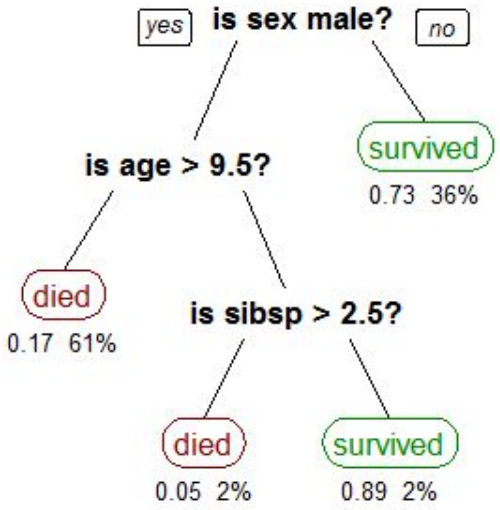
- POS n-grams
  - Word n-grams not good because prompt-specific
  - Features for n=1,2,3
- Transitional phrases
  - “also”, “as a result”, etc.
  - Measures flow between arguments
- Coreference
  - Measures how on-topic and unified the essay
  - Counts mentions of entities with each other

Topic Models

- LDA topic model: topic distribution in each essay is a mixture of several global topic distributions
- Our model trains a topic model using the set of essays for each score (one topic model for all essays with score=1, one for score=2,...)
- For each essay in the test set, compute log-likelihood that essay is generated by topic model
- Use log-likelihood as features in feature vector

Random Forests

- Based on decision tree classifier
- Random forest is made out of many decision trees
- Each decision tree made by picking random training examples and using a random subset of features to train the decision tree

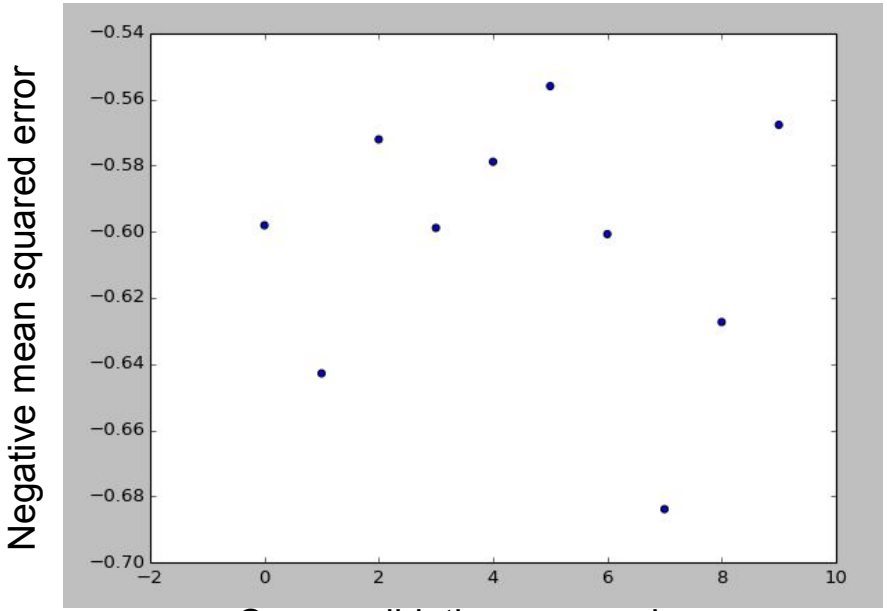


Results

	Accuracy	Mean deviation	Mean squared error	Pearson correlation
Baseline 1	0.44	0.67	0.93	0.00
Baseline 2	0.45	0.62	0.78	0.11
Persing and Ng	0.63	0.38	0.60	0.56

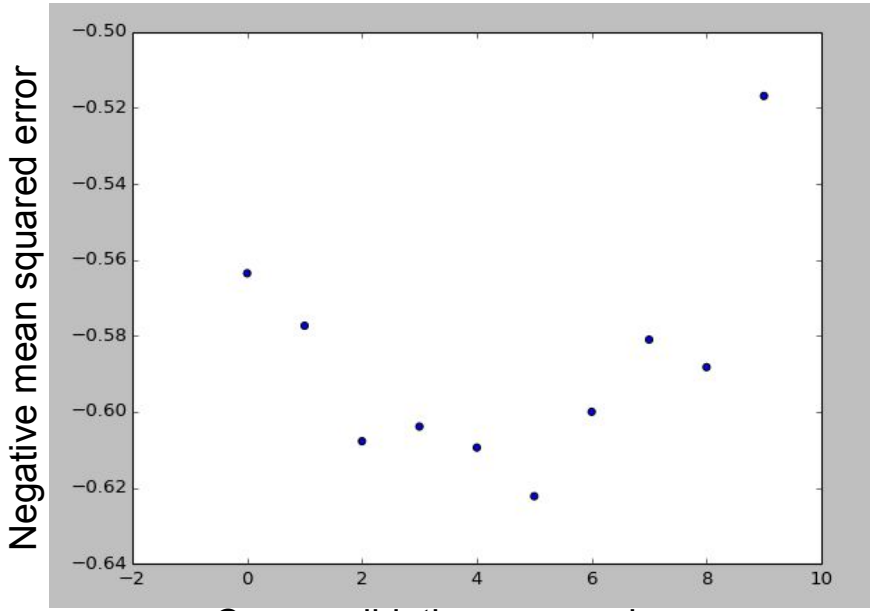
Results (cont.)

Baselines+POS+transitional phrases+coreference



mean: -0.599

LDA model+random forest classifier



mean: -0.587

Conclusions

- On average, our model performs slightly better than previous model
- Each category of features do not do well enough alone
- Future directions: try neural network, use other n-grams (at the phrase level, etc.)