USING HYBRID FEATURE IDENTIFICATION
& MACHINE LEARNING

AUTOMATIC ESSAY SCORING

AUTOMATIC ESSAY SCORING

- Use of specialized computer programs to assign grades to essays written in an educational setting.
- Classify a large set of textual entities into a small number of discrete categories => statistical classification.
- Significant factors: accountability, standards, technology, cost
- Improvement during the feature selection stage, rather than the training/testing stage.

DATASET: AUTOMATED STUDENT ASSESSMENT PRIZE

- The training and test data were acquired from a past competition from Kaggle.com sponsored by Hewitt-Packard.
- Each of the sets of essays was generated from a single prompt. Selected essays range from an average length of 150 to 550 words per response.
- All essays were hand graded and were double-scored. Each of the eight data sets has its own unique characteristics.

A TYPICAL DATAPOINT

- "Dear local newspaper, I think effects computers have on people are great learning skills/affects because they give us time to chat with friends/new people, helps us learn about the globe(astronomy) and keeps us out of **troble**! Thing about! **Dont** you think so? How would you feel if your teenager is always on the phone with friends! Do you ever time to chat with your friends or **buisness** partner about things. Well now - there's a new way to chat the computer, theirs plenty of sites on the internet to do so: @ORGANIZATION1, @ORGANIZATION2, @CAPS1, **facebook**, myspace **ect**. Just think now while your setting up meeting with your boss on the computer, your teenager is having fun on the phone not rushing to get off cause you want to use it. How did you learn about other **countrys**/states outside of yours? Well I have by computer/internet, it's a new way to learn about what going on in our time! You might think your child spends a lot of time on the computer, but ask them so question about the economy, sea floor spreading or even about the @DATE1's you'll be surprise at how much he/she knows. Believe it or not the computer is much interesting then in class all day reading out of books. If your child is home on your computer or at a local library, it's better than being out with friends being fresh, or being **perpressured** to doing something they know **isnt** right. You might not know where your child is, @CAPS2 **forbidde** in a hospital bed because of a drive-by. Rather than your child on the computer learning, chatting or just playing games, safe and sound in your home or community place. Now I hope you have reached a point to understand and agree with me, because computers can have great effects on you or child because it gives us time to chat with friends/new people, helps us learn about the globe and believe or not keeps us out of **troble**. Thank you for listening."

HISTORY OF AES

- Project Essay Grader, PEG (Ellis Page, 1968)
  - Proxy measures, including average word length, essay length, number of semicolons or commas.
  - Training stage and scoring stage.
  - Ignoring the semantic aspect of essays and focusing more on the surface structures.
  - NOT cost-effective.

HISTORY OF AES

- Intelligent Essay Assessor, IEA (Peter Foltz and Thomas Landauer, 1997)
  - Semantic text analysis, Latent Semantic Analysis (LSA)
    - meaning of word1+ meaning of word 2 + …+meaning of word n = meaning of passage
- It also measures grammar correctness and punctuation.
- Does NOT evaluate the creativity and reflective thinking.
- A product from Pearson Educational Technologies and used for scoring within a number of national exams.

ELECTRONIC ESSAY RATER (E-RATER)

- First used commercially in February 1999 by Jill Burstein in Educational Testing Service
- Corpus-based approach to model building, uses actual essay data to examine the sample essays.
- Employed in scoring the GMAT AWA since then. Test-taker's final score is determined through e-rater and one human-scorer.
- If there is discrepancy between e-rater and the human rater by more than 1 point, a second human rater in included.
- The discrepancy rate between e-rater and human raters has been less than 3%

SYNTACTIC MODULE

- A part-of-speech tagger is used to assign part-of-speech labels to all words in an essay.
- The syntactic “chunker” finds phrases and assembles the phrases into trees based on subcategorization information for verbs.
- Identify various clauses, including infinitive, complement, relative, and subordinate clauses and occurrences of modal verbs.

SYNTACTIC MODULE

- The number of complement, subordinate, infinitive, and relative clause and occurrences of modal verbs (would, could) to calculate ratios of these syntactic features per sentence and per essay ==> feature vector

- Limits of the Syntactic Module:
  - NOT enough syntactic features
  - Misspelling and wrong grammar

IMPROVING THE SYNTACTIC MODULE

- Importing more syntactic features
  - T-unit: one main clause with all subordinate clauses attached to it
  - mean length of clauses, mean length of sentences, mean length of T-unit clauses per sentence
  - clauses per T-unit, complex T-units per T-unit, dependent clauses per clause, dependent clauses per T-unit
  - complex nominals per clause, complex nominals per T-unit, verb phrases per T-unit

Table 4. Correlations between complexity scores computed by the two annotators

Measure	Correlation	Measure	Correlation
MLC	.985	DC/T	.981
MLS	1.000	CP/C	.964
MLT	.998	CP/T	.965
C/S	.978	T/S	.969
C/T	.978	CN/C	.948
CT/T	.912	CN/T	.957
DC/C	.954	VP/T	.958

Corpus: Written English Corpus of Chinese Learners

This corpus comprises 3,554 essays written by English majors from nine different four-year colleges in China.

These essays contain an average of 315 words, with a standard deviation of 87.

IMPROVING THE SYNTACTIC MODULE

- Improving the tolerance of the parser with penalty
  - If there is a word that is a OOV, tag it and penalize for the misspelling.
  - If the POS tags cannot be parsed into a tree, try to match it with a near-Tree and penalize for the difference.
  - Same/similar errors should be penalized for only once.

RHETORICAL STRUCTURE ANALYSIS

- Use rhetorical cue words and structure features, in addition to other topical and syntactic information to identify discourses.
- Automated Argument Partitioning and Annotation
  - The argument units of the essays are labeled as “marking the beginning of an argument”/“marking argument development”
  - These argument partitioned version of essays are used by the topical analysis module to evaluate the content of individual arguments.

RHETORICAL STRUCTURE ANALYSIS

- Limits of the Automated Argument Partitioning and Annotation system:
  - Rely heavily on the cue words
  - Argument phases are only classified into 2 stages
  - Arguments are considered in paragraphs rather than sentences

IMPROVING THE DISCOURSE MODULE

- Use discourse indicators (similar as the cue words) to classify each clause as major claim/claim/premise/non-argumentative
  - Discourse indicators: support/conflict relation types, as well as domain terminology
- Use argumentation scheme structure to improve the partitioning. Fit the clauses into a existing schema by their categories

Expert Opinion
Premise: Source E is an expert in subject domain S containing proposition A [FieldExpertise]
Premise: E asserts that proposition A is true (false) [KnowledgeAssertion]
Conclusion: A is true (false) [KnowledgePosition]

Positive Consequences
Premise: If A is brought about, then good consequences will (may plausibly) occur [PositiveConsequences]
Conclusion: Therefore, A should be brought about [EncouragedAction]

Table 4: Argumentation schemes

Relation Type	**Words**
Support	because, therefore, after, for, since, when, assuming, so, accordingly, thus, hence, then, consequently
Conflict	however, but, though, except, not, never, no, whereas, nonetheless, yet, despite

Table 1: Discourse indicators used to determine propositional connections

TOPICAL ANALYSIS

- Evaluates the lexical and topical content of an essay by comparing the words it contains to the words found in manually graded training examples
- Two programs: one based on word frequency (EssayContent) and the other on word weight (ArgContent)

- EssayContent:
  - convert the vocabulary of each score category to a single vector whose elements represent the total frequency of each words
  - computes the cosine similarity between the vectors
    - NOT sensitive to essay length

TOPICAL ANALYSIS

- ArgContent assign scores to each argument generated by the Discourse Analysis Module
- The word frequency vectors are converted to vectors of word weights. The weight for word i in score category s is:

$$w_{i.s} = (freq_{i.s} / max\_freq_s) * \log(n\_essays_{total} / n\_essays_i)$$

- Each argument is evaluated using cosine similarity and uses the adjusted mean of the set of scores.

TOPICAL ANALYSIS

- Limits of Topical Analysis:
  - Does NOT check the line of reasoning and logic of the argument
  - Does NOT take the repeated usage of certain words into consideration when calculating the weights
  - Less weight does not necessarily mean “bad word choice“, should adjust weight according to some standards

IMPROVING THE TOPICAL ANALYSIS

- When comparing the similarity between arguments, should also compare the similarity between corresponding argument components (premises, conclusion)
- Use the Argumentation Scheme Structure developed in Discourse Analysis Module to score the reasoning of arguments
- Adjust the word weights using information from domain terminology, synonyms, and word usage traditions

TRAINING AND TESTING

- Syntactic, discourse and topical analyses yielded a total of 57 features for each essay. By improvement, there would be far more features for training/testing.
- A stepwise linear regression analysis was used to compute the optimal weights for these features during training stage.
- The criteria differs by topic.
- We don’t know which ML algorithm is used by ETS currently to do the training/testing process of AES.

REFRENCES

- Jill Burstein. 2003. The e-rater scoring engine: automated essay scoring with natural language processing. Mark D. Shermis and Jill C. Burstein (Eds.). Automated essay scoring: a cross disciplinary approach.
- Jill Burstein, Karen Kukich, Susanne Wolff, Chi Lu, Martin Chodorow, Lisa Braden-Harder, Mary Dee Harris. 1998. Automated scoring using a hybrid feature identification technique. In the Proceedings of the 17th international conference on Computational linguistics - Volume 1.
- John Lawrence, Chris Reed. 2015. Combining Argument Mining Techniques. In the Proceedings of the 2nd Workshop on Argumentation Mining, pages 127-136.
- Semire Dikli. 2006. Automated Essay Scoring. Turkish Online Journal of Distance Education - Volume 7, Issue 1.
- Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing.