iLit
inspireLiteracy

ABOUT PEARSON KNOWLEDGE TECHNOLOGIES INTELLIGENT ESSAY SCORING

PEARSON

Pearson Knowledge Technology writing software is powered by the Intelligent Essay Assessor using scores assigned by human raters to several hundred representative student essays all written in response to a particular essay prompt or question for a particular grade level. By using computational modeling, IEA mimics the way in which human readers score. In study after study comparing the performance of IEA to that of skilled human graders, the quality of IEA's assessment equals or surpasses that of the humans.

First, a set of representative student essays are collected and scored independently by two or more human graders. Usually 200 to 250 doubly scored human papers are sufficient. A regression model with about 50 content and computational linguistic variables is used to predict the average human score. A separate regression model is calculated for each essay prompt.

By far the most important variable for matching human scores turns out to be the essay's content. This variable uses Latent Semantic Analysis (LSA). Latent Semantic Analysis is a computer model that was invented and patented by several Knowledge Technologies employees in the late 1980s and is now in wide use around the world. LSA automatically constructs a semantic space (a number representing the meaning of each word) by analyzing large volumes of text that an average student would encounter and read through high school. The text corpus for this includes all the paragraphs from about 12 million running words of text. LSA uses as input a co- occurrence matrix of words and their frequency in paragraph units. This input matrix is reduced to one of much smaller rank, using Singular Value Decomposition (SVD), a matrix algebra technique similar to factor analysis. SVD is a least squares approximation of the original matrix. It usually uses 300 independent vectors to represent each word and each paragraph in the text collection. In the end, the analysis assures that every paragraph is the sum of the 300 element vectors for its words, and every word is the average ofall the vectors standing for the paragraph that uses the same vocabulary corpus, not just those already in the corpus. A variety of analyses and applications have found that LSA usually agrees with the human judgments of the similarity of two paragraphs or words 90 percent as well as two humans agree with each other.

For scoring an essay, the 200 to 250 training essays are each given a 300-dimensional score by averaging the word vectors occurring in each essay. That is, each word is represented by a vector with 300 real numbers correspondingto each of the dimensions — the separately measured quantities describing the essay. New essays to be graded are given a 300 dimensional score using the words that occur in them and averaged over each of the 300 dimensions.

Next, the new essay is compared to each of the training essays in terms of similarity (cosine of the angle between the two essays or Euclidean distance between the two). The closest neighbors to the new essay and training essays determine the content score. Essays with high scores will tend to cluster. So, a new essay close to high scoring training essays will receive a high score. Off-topic essays can be flagged automatically because they have insufficient content similarity to the training papers.

Many other automatically (thus consistently) used variables are also used to score each essay to insure that factors not captured by LSA are not ignored. Virtually all the separate characteristics of student essays on which teachers base grades, comments and corrections influence PKT scores to approximately the same extent that they do for human scorers. This is also true of the characteristics described in the rubrics that human graders seek to follow. Measures based on the raw length of essays, sentences or paragraphs are never used. Similarly, keywords, such as ones that signal an essay's organization (e.g. "first," "in conclusion," "thus," etc.) are not given special weight. These types of variables are too highly coachable. If it were known that using them increased scores, beating an automatic essay grader would be quite simple. A separate regression model is calculated for each essay prompt.

A prompt independent grading model has also been developed that will score an arbitrary essay based only on the grade level of the student. Because the scoring engine is not trained on essays responding to a particular prompt, the scoring is based on stylistic, grammar, usage, and mechanics variables. The scoring engine has no way of factoring in the content of the essay. However, it is easier and less expensive to usethe prompt independent model.

While a bit of accuracy is sacrificed — a decrease of ~0.1 in the reliability coefficient — it is easy for teachers to customize the prompts to their lesson plans. The downside of the prompt independent method is that the score uses only linguistic, stylistic, vocabulary, and mechanics variables.

iLit

inspireLiteracy

PEARSON

redefiningliteracy.com