Bull-O-Meter: Predicting the Quality of Natural Language Responses

Markus Krause
UCBerkeley, ICSI
public.markus.krause@gmail.com

Abstract

Grading essays in large online classes or predicting the quality of a text submission in a crowdsourcing task is challenging. In this paper we propose a new natural language model to predict the quality of natural language responses.

Introduction

With the rise of MOOCs automated grading of essays becomes a popular topic again (Balfour, 2013). A tool able to predict the quality of an essay can also be helpful in predicting the response quality of a natural language task in crowdsourcing. Using crowdsourcing to generate natural language data is a common practice (Aras, Krause, Haller, & Malaka, 2010) yet ensuring data quality especially for more complex tasks is challenging. In this paper we propose a new natural language model to train regression algorithms to predict the perceived quality of natural language responses. We compare our model to nine state of the art essay scoring engines. The comparison is done using eight data sets of student essays from six US States. Each sample consists of text and ratings from at least two human raters along with a final score. The essays encompassed writing assessment items from three grade levels (7, 8, 10). Six of the eight essay sets were transcribed from their original handwritten responses using two transcription vendors. The data set was made public by the *Hewlett Foundation* as part of their *Kaggle* challenge on *Automated Essay Scoring*. Table 1 gives an overview of the 8 essay sets. To compare systems we use Pearson's *r*. We compare our model to the results of nine other scoring engines as reported by Shermis and Hammer (Shermis & Hamner, 2012). Our model consists of 113 features. We extract a feature vector for each essay in a training set. Each essay has a number of associated human ratings. We generate a vector for each rating and one vector for the average of all human raters. For the first essay set with 1785 entries our training set consists of 5,355 samples. Our Random Forrest Regressor (Liaw & Wiener, 2002) generated 100 random trees per forest using mean squared error as split criterion. We calculate Pearson's *r* between the average human rating and our models prediction for the separate test sets.

Essay Set	*Train Samples*	*Test Samples*	*Average Words*	*Ratings per Essay*
1	1785	589	360	2
2a	1800	600	380	2
2b	1800	600	380	2
3	1726	568	110	2
4	1772	586	94	2
5	1805	601	122	2
6	1800	600	153	2
7	1730	495	171	8
8	918	304	622	12

Table 1. Basic statistics of the eight data sets used.

Language Model

We base our linguistic model on a feature set that has previously been used to investigate writing styles in educational settings (Kilian, Krause, Runge, & Smeddinck, 2012; Krause, 2014). We use the following set of features: length frequencies (word length, sentence length), emotional content (valence and arousal), language specificity frequency, part of speech frequency, and sentence mood. We preprocessed all reviews with the NLTK part-of-speech (POS) tagger (Bird, Klein, & Loper, 2009). We then filtered stop words and words not in Wordnet (Miller, 1995). Wordnet is a natural language tool that provides linguistic information on more than 170,000 words in the English language. We also lemmatized the remaining words to account for different inflections.

Part of Speech Tag Frequency: For this feature set we use the Penn Treebank part of speech tag set. We use *pattern.en* to extract these tags. We calculate the relative frequency of each tag. Giving a total of 35 features.

Text length: the first two feature sets we use the frequency of number of letters in words and the frequency of number of words per sentence. For word length frequency we considered only those words that have a *Wordnet* entry and are

	H1H2	AIR	CMU	CTB	ETS	MI	MM	PKT	PM	VL	BoM
1	0.73	0.8	0.79	0.71	0.82	0.82	0.66	0.8	0.76	0.8	0.87
2a	0.8	0.68	0.71	0.69	0.74	0.72	0.62	0.7	0.72	0.71	0.9
2b	0.76	0.67	0.64	0.64	0.7	0.71	0.55	0.65	0.69	0.69	0.84
3	0.77	0.72	0.74	0.69	0.72	0.75	0.65	0.66	0.73	0.73	0.83
4	0.85	0.76	0.81	0.76	0.82	0.82	0.68	0.75	0.76	0.8	0.92
5	0.75	0.82	0.81	0.8	0.81	0.84	0.65	0.8	0.78	0.83	0.81
6	0.74	0.76	0.77	0.65	0.77	0.81	0.66	0.75	0.78	0.77	0.81
7	0.72	0.71	0.78	0.75	0.81	0.84	0.58	0.78	0.8	0.82	0.79
8	0.61	0.71	0.66	0.63	0.71	0.73	0.62	0.7	0.68	0.72	0.74
	0.748	0.74	0.75	0.7	0.77	0.78	0.63	0.73	0.74	0.76	0.83

Table 1: Pearson correlation between the nine different grading methods and the average score of human raters on the eight data sets. ***Bull-o-Meter(BoM)****, Vantage Learning (VL), American Institutes for Research (AIR), Measurement, Inc. (MI), TELEDIA, Carnegie Mellon University (CMU), MetaMetrics (MM), McGraw-Hill (CTB), Educational Testing Service (ETS), H1H2 refers to the average agreement between two human raters. The last row shows the average Pearson correlation.*

not stop words. Furthermore we group words longer than 20 characters in one group so word length frequency gives us 20 features. The sentence length was measured including all words returned by the POS-tagger. We grouped sentences longer than 30 words into one group, so sentence length frequency gives us 30 individual features.

Emotionality: The next two feature sets we looked at were valence and arousal. Valence refers to whether a text is positive, negative, or neutral, and arousal represents how strong the valence is. We use 5 levels of arousal and valence both ranging from 1 to 5 so 10 features total. We used *pattern.en*, a tool based on *NLTK*, to extract valence and arousal.

Specificity: Another feature set we explored was specificity, which refers to how specific the words in a text are. We measured specificity by determining how deep each word appears in the *Wordnet* structure. Words that are closer to the root are more general (e.g. *dog*) and words deeper in the *Wordnet* structure are more specific (e.g.). Word depth ranges from 1 to 20 (20=most specific).

Sentence Mood: the last features we considered involves looking at moods of sentences. Each sentence was classified as either indicative (written as if stating a fact), imperative (expressing a command or suggestion), or subjunctive (exploring hypothetical situations). We again used *pattern.en* to extract sentence mood.

Results

As Table 1 shows our model consistently outperforms other essay grading engines on almost every data set in our experiment. Figure 1 also illustrates the consistency of our model even for scores with minimal training data. As we only report initial results we do not yet have a detailed analysis of feature importance. However, word specificity

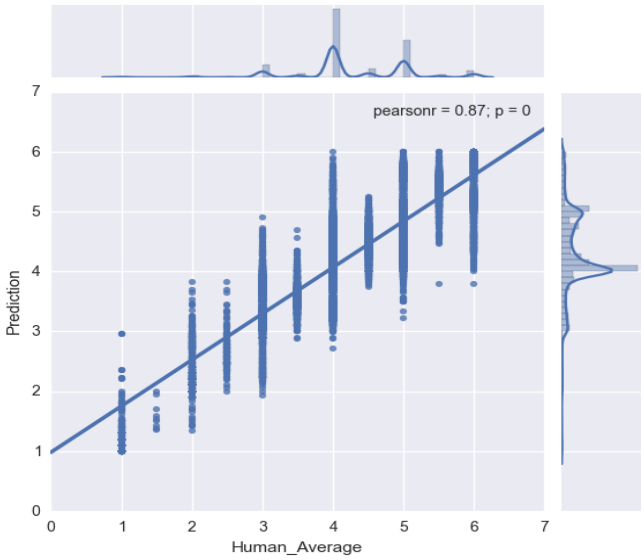

Figure 1: Correlation of our model and two human raters on the first data set of the Automated Essay Scoring challenge.

and sentence mood seem to have a positive effect and are also rarely used in other grading engines. Emotionality of a text seem to affect the results positively. We speculate that the reason for this effect is subjective nature of grading essays. Reviewers might prefer a certain style. This style includes how positive or negative an essay is written. It remains an open question if this also applies to the prediction of response quality in crowdsourcing tasks.

References

Aras, H., Krause, M., Haller, A., & Malaka, R. (2010). Webpardy: Harvesting QA by HC. In *HComp'10 Proceedings of the ACM SIGKDD Workshop on Human Computation* (pp. 49–53). New York, New York, USA: ACM Press.

Balfour, S. (2013). Assessing writing in MOOCS: Automated essay scoring and Calibrated Peer Review. *Research & Practice in Assessment*, *8*, 40–48.

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*.

Kilian, N., Krause, M., Runge, N., & Smeddinck, J. (2012). Predicting Crowd-based Translation Quality with Language-independent Feature Vectors. In *HComp'12 Proceedings of the AAAI Workshop on Human Computation* (pp. 114–115). Toronto, ON, Canada: AAAI Press.

Krause, M. (2014). A behavioral biometrics based authentication method for MOOC's that is robust against imitation attempts. In *Proceedings of the first ACM conference on Learning @ scale conference - L@S '14* (pp. 201–202). Atlanta, GA, USA: ACM Press.

Liaw, A., & Wiener, M. (2002). Classification and Regression by Random Forest. *R News*, *2*, 18–22.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39–41.

Shermis, M. D., & Hamner, B. (2012). *Contrasting State-of-Art Automated Scoring of Essays: Analysis*.