Grading emails and generating feedback

Abhishek Unnam, Rohit Takhar, Varun Aggarwal
Aspiring Minds
{abhishek.unnam, rohit.takhar, varun}@aspiringminds.com

ABSTRACT

Email has become the most preferred form of business communication. Writing *good* email has become an essential skill required in the industry. *Good* email writing not only facilitates clear communication, but also makes a positive impression on the recipient, whether it be one's colleague or a customer. The aim of this paper is to demystify the components of a *good* email and to define a set of parameters by which to grade the quality of an email and provide detailed feedback. This can help candidates improve their email writing skills and also guide tutors. The email grading parameters encompass traditional attributes of written English (i.e. coherent and relevant content and correct grammar) but also include a unique set of characteristics that we may objectively identify as email etiquette. These characterstics comprise the metrics we use to evaluate the quality of the various constituent parts of an email. We grade the email using artificial intelligence, acting on semi-structured text. We use a mix of machine learning and rule-based systems to effectively grade an email on the specified parameters. Our system automatically grades email with accuracy comparable to human graders.

Keywords

Automatic grading; Email Writing; Semi-Structured Text analysis; Education; Supervised Learning

1. INTRODUCTION

In today's knowledge economy, good communication skills (spoken as well as written) are vital for success in the workplace. According to the O*NET taxonomy of jobs and skills [5], 53% of all jobs require a moderate to high level of writing and speaking skills. Lately, there has been a lot of work [18], [10], [17] on automated evaluation of speaking skills. Some of these automated systems help candidates to improve their language-speaking skills [3], [7]. The evaluation of writing skills is generally thought to be confined to academic essay grading [1], [14]. These studies were prompted by the demand for more efficient evaluation of high-volume educational/academic tests such as TOEFL and SAT. There are job tasks that mimic essay writing such as writing user manuals, product documents or filing Request-for-Proposals (RFPs). However, the writing that is most ubiquitous in companies is email. Interactions with clients and all manner of internal communications with managers, peers, and support services (i.e. human resources, tech support, etc) are carried out via email. Therefore, composing *good* email has become a necessary skill.

We wish to automatically grade email writing skills and provide feedback. This will create a mechanism for students and jobseekers to get feedback on their email writing skills and also provide companies with the ability to automatically test email writing skills of job candidates on a large scale and use the grades in the hiring process. We wish to automatically grade email writing skills and provide feedback. This will create a mechanism for students and jobseekers to get feedback on their email writing skills and outline a path for improvement. It will also provide companies with the ability to automatically test email writing skills of job candidates on a large scale and use the grades in the hiring process. Despite the importance of email writing skills, we have not found any previous work in this area.

We first demystify the components of a *good* email. The grading parameters consist of the traditional attributes of written English (i.e. coherent and relevant content and correct grammar) but also include a unique set of characterstics that we objectively identify as email etiquette. We have identified a set of 36 metrics on which to evaluate the quality of the various constituent parts of an email. Many of these metrics are derived from the general norms of communication, while others are specific to the written form of email (discussed in detail in Section 2). Our aim is to automatically grade email on each of these metrics.

We grade email using artificial intelligence, acting on semi-structured text. Some parts of an email are structured, such as the subject, salutation and sign-off (Figure 1). The body of the email, a set of sentences, is unstructured. There are multiple approaches to grading structured and unstructured text. Essay grading [4] and the grading of short answers [13] are examples of unstructured text grading. Generally, researchers generate a set of features such as bag of words, word embeddings, parts-of-speech (POS) tags, etc. and use supervised or semi-supervised learning to predict grades. The grading of computer programs which use tightly defined grammar is an example of structured input grading. In [15] authors derive sophisticated features by exploiting the

Abhishek Unnam, Rohit Takhar and Varun Aggarwal "Grading emails and generating feedback" In: *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019),* Collin F. Lynch, Agathe Merceron, Michel Desmarais, & Roger Nkambou (eds.) 2019, pp. 444 - 449

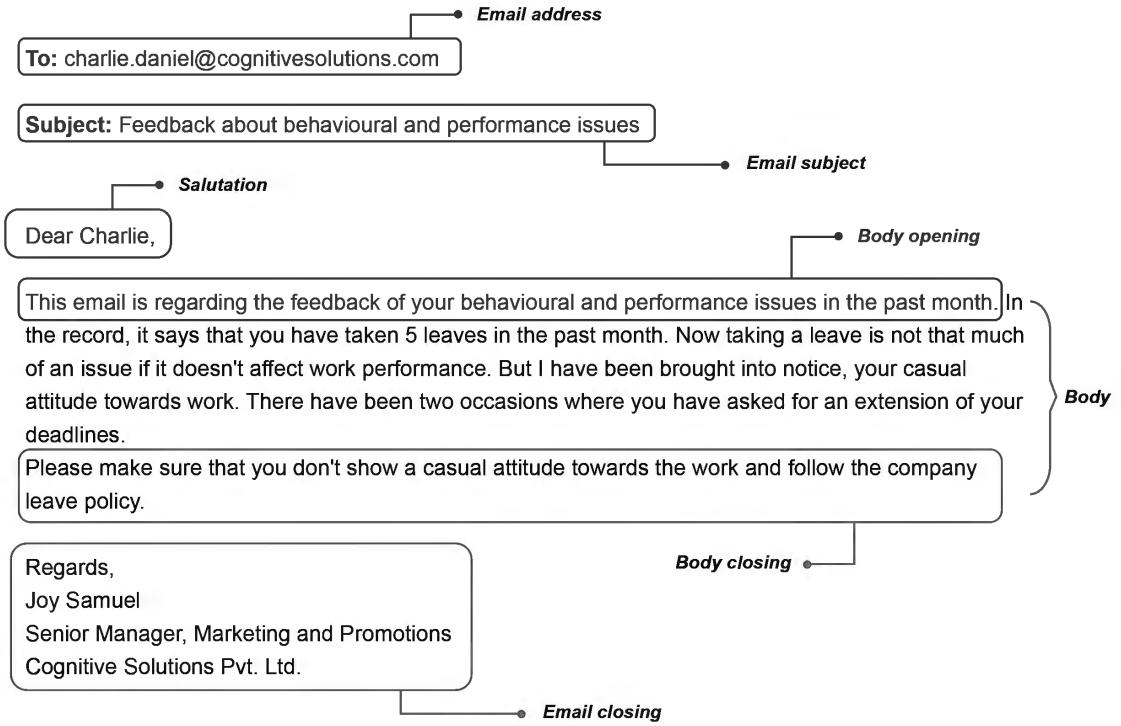

Figure 1: Sample email being written by a student.

structure of the program by drawing control row and data dependency graphs. Whereas, resume parsing is an example of processing semi-structured text [2].

We propose a mix of machine learning and rule-based models to grade the various quality metrics of an email. We find that such an approach works well for semi-structured text and for a variety of evaluation parameters. We show that generally the rule-based models work well for processing the structured parts of email and for grading email on etiquette. On the other hand, content relevance and grammar are better graded via machine learning. Interestingly, for some attributes, a mix of both models work best. Our approach stands contrast to recent trends that apply machine learning to all tasks indiscriminately.

We derive a number of natural language features for our machine learning models. We use supervised learning to build the models. The rule-based models use word lists and regular expressions. Our algorithm provides accuracy that rivals expert consensus. Our final system generates separate grades for content relevance, grammar and email etiquette. It also provides detailed feedback on the types of errors that occur in each part of the email. In particular, the paper makes the following contributions:

- We propose the first viable system for automatically grading email writing skills and delivering constructive feedback.
- We demystify the components of *good* email and provide objective evaluation criteria.
- We propose a mix of machine learning and rule-based models for evaluating emails. The accuracy of our system rivals expert consensus.
- We provide first-of-its-kind data-based insight into the errors committed by students and jobseekers.

The paper is organized as follows. In Section 2, we describe the metrics we define for evaluating email etiquette. In Section 3, we discuss the methods used to evaluate the metrics. In Section 4 we explain the experiments, datasets and models that have been developed for grading. In Section 5 we present our results by comparing the performance of our system with that of human experts. Concluding remarks are provided in Section 6.

Response	Status
I care for this in an email that I receive	agree
I may care for this in an email that I receive	not sure
I do not care for this in an email that I receive	disagree

Table 1: Each rule was put into one of these categories

2. EMAIL ETIQUETTE

Quality metrics for evaluating email writing skills go beyond the parameters traditionally defined in the grading of written language, i.e. paragraphs and essays.

Email as social communication must follow certain norms. Some of these norms are derived from the general norms of verbal communication, while others are specific to the written form of email. These norms manifest themselves as rule sets for the structured fields of email and the unstructured body. A simple rule is that values for all the structured fields should be present. For example, an email that lacks a subject line is

Quality Metrics	Explanation & Examples	Counts
Missing	Missing subject line, salutation, signoff etc. **Subject:** **Email Body:** This is to inform you that....	3
Redundancy	Starting the subject line with terms such as 'regarding', 'response to'. **Subject:** regarding behavioral and performance issues.	6
Word usage	Incorrect usage of words in various sections of an email - using names/greetings in subject line, usage of informal, abbreviated words etc. **Subject:**Employee feedback for Charlie Daniels **Email Body:**Hi Daniel, Can u pls respond quickly...	10
Style	Errors specific to conventions like greeting and sign-off style. **Email Body:** Hella Daniel/Heyy Charlie Daniels/Hi Mr Charlie Daniel ..... Yours Sincerely/Truly/faithfully Mr Charlie Daniel	7
Emotional Punctuation	Errors like using too many commas inside a sentence, using exclamation/semi-colons marks inside subject/salutation/closing, using all uppercase words in subject line etc. **Subject: POOR performance !!! neeed improvement** **Email Body:** Heyy!! Charlie, HI CHARLIE please reply...Thanks, Alisha	5
Punctuation	Capitalisation errors like starting subject line with lowercase, proper nouns starting in lowercase etc. Not giving space after fullstop. **Subject:** feedback on performance **Email Body:** hi charlie daniel, This is to inform you about the poor performance in last financial year.I have seen many instances of work lapse.	5

Table 2: Quality metrics with explanations & examples.

not good. A rule derived from verbal communication norms is that the opening greeting should not address the recipient by both first and last name, and that a title should be added when addressing a person by last name only. An example of a rule typical to the written form is that the subject line should not be longer than a given length. Another example is that no words (except acronyms) should appear in all upper case, and that emoticons should be avoided. There are also other metrics that derive from the changing norms of communication. In today's business world, one doesn't address others as 'Respected' and rather use a 'Hi/Hello/Dear'. There are similar rules that govern the body of an email.
There are additional regarding the purpose of the email, or in a grading scenario, the prompt of the email writing task. For example, when responding to an irate customer, one should not employ the oft-used email phrase "Hope you are doing well". Similarly, the closing line of an email may depend on the prompt. Emails that seek a response may conclude with, "I look forward to your response", while a simple conversation may end with simply "Feel free to reach out to me".
There is no standard list of these rules. We looked at the various blogs and documents with email writing rules to put together a super-set of 57 different possible rules. We shared these rules with three professionals from India and three from US, each with 10+ years experience in the corporate world. Two from each country were from either the IT, banking or contact center industry. Three were involved in external communications, while another three were into internal communications. We asked them to rate each rule as in Table 1. From the original 57 rules, we selected the 36 on which four

Section	Number of errors
Subject	9
Salutation	6
Email Body	13
Closing	8

Table 3: Section-wise error counts.

or more of the professionals agreed. These 36 rules became our quality metrics, which we categorized into 6 broad categories with definitions (See Table 2). In Table 3, we show the number of rules that apply to each part of an email. We evaluate an email against each of these rules and delineate the errors in a candidate feedback section. We also provide a total score for email etiquette. This score is based on the number and severity of errors made. For example, a *missing* error is more severe than *redundancy* errors. We also include the standard parameters and rubrics [8] of English evaluation: content relevance and grammar.

3. METHODS USED

We used three different methods to evaluate the various facets of quality metrics. We use a mix of rule-based and machine learning methods.

3.1 Word list-based

Many email etiquette errors may be detected simply by presence or absence of certain words or phrases. For instance, an email should not have abusive words or slangs. On the other hand, a pleasing email shall have words like 'thank you', 'please', 'request' etc. We created word-lists to evaluate such metrics in an email. Such an approach has also been followed earlier in emotion detection [16]. The words in email are matched against the word lists after stemming. Based on the occurrence and counts (normalized, in some cases), the feedback and scores, are respectively generated. For instance, if someone writes "Whoa, It was great working with u." we provide a feedback to avoid the usage of informal words like 'whoa' and 'u'. We graded 19 error metrics in this manner using 26 word lists[1]. The efficacy of this approach was tested by evaluating the generated feedback and score against human expert feedback. This is described in Section 4.

3.2 Pattern-based

Certain metrics were evaluated not by the occurrence of a single word or phrase, but based on an expected pattern. For instance, the recipient may be addressed by using more than one combination of the greeting (Dear/Hi/Hey), title (Mr/Ms/Dr), first name and last name. Some combinations are right (Dear First Name), while some are wrong (Dear Last Name). The correct and incorrect patterns were coded into regular expressions, to provide exact detection of an error and providing feedback (e.g., 'Do not use only last name without title'). One may note that this detection assumes we know the full name of the recipient. For every email writing task (prompt), certain structured information such as recipient name, company name, sender name and certain keywords is generated at the beginning (manually). This is used for various error detection including, for example, capitalization, and spelling errors. In total, 16 error metrics were evaluated using pattern matching.

3.3 Regression-based

We use supervised learning to grade an email on content relevance and grammar. We train regression models using a number of different features to predict expert grades. We now describe the features used for each of the two parameters.

3.3.1 Content Relevance

The candidate's task was to write an email according to the situation provided in the prompt. We wish to evaluate whether the content properly addresses the situation, is comprehensive, coherent and without unnecessary information. This metric is linked to the semantics of the email only and doesn't evaluate other parameters such as the emotion of the email. We describe the natural language features used for the task.

- **Word embeddings**: Here, we used the Word2vec model [11], particularly Google's pre-trained model developed with a vocabulary over 3 million words and phrases and trained on roughly 100 billion words from Google's news dataset. For each word in the email we first calculate a 300-length lower dimensional vector and then sum it across all the words in the email.
- **Bag of Words (BOW):** We used the bag of words feature-counts of unigrams, bigrams and trigrams. All the words were stemmed and stops words were removed.
- **Prompt Overlap:** We calculated prompt overlap in two ways: *exact match* and *extended match*. In *exact match*, we count the number of common words between the prompt and the email. In *extended match*, we add the synonyms of all words in the prompt using WordNet [12]. We then count the number of common words between the extended prompt word list and the words in the email.

3.3.2 Grammar

Below are the features we used to evaluate the grammatical correctness of Here our aim is to evaluate the grammatical correctness of the email. We use the following features:

- **Bag of POS tags:** Here, words are assigned to their respective part of speech (POS) tags using the Penn Treebank NLTK tagger [9]. We then considered bigrams and trigrams of POS tags. This feature removes the semantic information from the words, while preserving the sentence structure and grammatical features.
- **Error Counts:** We also use counts of the grammatical errors in the email as identified by open-source grammar correction tools.
- **Proportion of good tags:** Here, we wish to find the similarity of the language in the email with that of a grammatically correct corpus. We used Brown corpus [6] for our purpose. We generated bag of POS bigrams and trigrams from this corpus. We consider the top 70% most frequently occuring POS bigrams and trigrams as a set of *good* n-grams. We then find what proportion of n-grams in the email are *good* n-grams.

4. EXPERIMENTS

We had designed our experiments to answer the following questions:

- How accurate is our approach in predicting content and grammar scores as compared to human experts?
- How accurately do we detect email etiquette errors using the word list and the pattern matching methods?
- What proportion of errors marked by humans experts do we correctly detect and how many false errors do we generate?

We conducted our experiments on a set of 1200 emails which were manually rated by human experts. For training models, we made use of both linear (linear, ridge regression) and non-linear (random forest(R.Forest), SVM) techniques. We discuss more about the dataset in the next section.

4.1 Dataset

Our dataset consists of 1200 emails in response to three different prompts. We used an equal set of 400 emails per prompt, after removing any blank email or those with very little content . These samples were collected from both undergraduate and graduate students. The candidates were given a situation and asked to write an email to address the situation. The three situations included a customer service situation where one needs to address a customer's complaints, a sales situation where one probes the requirement of a prospect and

[1]For some error metrics more than one word lists were used.

		Prompt 1		Prompt 2		Prompt 3	
Model	**#Features**	**Train (r)**	**Validation (r)**	**Train (r)**	**Validation (r)**	**Train (r)**	**Validation (r)**
Linear	50	0.86	0.79	0.87	0.81	0.85	0.77
Ridge	50	0.86	0.80	0.87	0.81	0.84	0.77
R.Forest	150	0.95	0.80	0.95	0.82	0.92	0.79
SVM	50	0.84	0.79	0.86	0.80	0.83	0.76

Table 4: Performance of prompt-specific content models.

			Validation			
Model	**#Features**	**Train (r)**	**Overall (r)**	**Prompt 1 (r)**	**Prompt 2 (r)**	**Prompt 3 (r)**
Linear	75	0.77	0.62	0.59	0.65	0.65
Ridge	75	0.70	0.66	0.62	0.67	0.75
R.Forest	150	0.85	0.73	0.71	0.73	0.74
SVM	150	0.62	0.60	0.52	0.59	0.74

Table 5: Performance of grammar models trained on complete set of emails. We present the overall and prompt wise validation correlations(r).

promotes a service and a people management situation where one needs to give feedback to an employee on performance issues. The responses were collected using $AMCAT$[2], our proprietary computer based testing platform.

All email responses were graded by human experts. Content Relevance and Grammar were graded based on a 7 point and 5 point rubric respectively. Detailed guidelines were provided for identifying each email etiquette error. The experts had to mark the exact location where the error occurred and the category of the error. Each email was graded by 3 different experts. These included an English language trainer, a sales manager and a customer service manager. Each of these had an experience of more than seven years in the industry. The experts first went through a three-day training where they learned how to interpret the rubric and were subjected to practice grading exercises.

We achieved an average inter-rater correlation 0.83 for content scores and 0.74 for the grammar scores. For email etiquette, only those errors were considered where atleast two experts had a consensus. A consensus was reached for 83% of the total cases.

4.2 Models

For content relevance we trained different models for each prompt, while for grammar, a generic model was trained across all prompts. For each model the corresponding dataset was divided into train and validation sets. We used a stratified 70-30 split for train-validation sets. We used linear regression, linear regression with L2 regularization (ridge), SVM and R.Forest to train models. Select K-best algorithm was used for feature selection. The models with the lowest cross-validation (4-fold) error were selected.

For the rule-based system, we used 50% of the email sets to experiment with the rules, patterns and word-lists. Several iterations were performed to fine tune the algorithm. The rules were then tested on the remaining 50% data. We report the results on this validation set.

[2]https://www.aspiringminds.com/contentTech

5. RESULTS

We evaluate our machine learning models using the Pearson correlation coefficient between predicted and the expert grades. We report and discuss results for the validation set. For content evaluation, all modeling techniques provide similar results. The correlation for all prompts is around 0.79 (refer Table 4). For grammar scores, $R.Forest$ gives the best results, though marginally better than other techniques. Here, we created one model across prompts and the overall correlation across prompts is 0.73. Also, the correlation value of any individual prompt is more than 0.71 (refer Table 5). For email etiquette errors, we report two metrics of errors:

Error Category	**Counts(%)**	**TP(%)**	**FP(%)**
Redundancy	31.27	70.00	17.86
Missing	69.77	93.55	11.29
Punctuation	90.78	94.12	19.17
Emotional Punctuation	18.00	84.48	6.90
Style	95.74	95.11	8.27
Word usage	17.05	100.00	20.00
Average	53.76	89.54	13.91

Table 6: Category wise performance of rule based system for email etiquette.

- **TP (True Positives)**: It is the proportion of expert errors that were correctly identified. An error is deemed correctly identified only when the position and error type is correct.
- **FP (False Positives)**: It is the number of non-existent errors identified, normalized by the total count of expert errors. This provides an idea of the proportion of extra errors detected.

We report all the values for the unseen validation set, 50% of the total data. We present the category-wise and section-wise results in Table 6 and Table 7 respectively. **Counts(%)** (in Table 6 and 7) states the percentage of emails that had the given error. This provides some interesting insights. For

Section	Counts(%)	TP(%)	FP(%)
Subject	70.41	70.59	8.82
Salutation	79.84	92.78	17.53
Closing	94.88	95.28	11.02
Email Body	77.43	89.83	10.10

Table 7: Section wise performance of rule based system for email etiquette.

instance, we find that most students make an error in the email closing, followed by salutation and the least number of errors in writing the subject line. On the other hand, the most number of emails are impacted by errors in style and classical punctuations. When we analysed in detail, we find that the *style* of the *closing* has the most number of errors. Candidates either completely miss the sign-off or use overly formal phrases like 'Yours Truly/Sincerely/faithfully'. We find that the average TP is 89.54% and FP is 13.91%. One may recall that the expert consensus was achieved in 83% of the cases. Our system detects most of the expert-identified errors and has a low rate of detecting false errors. Further, we find that the lowest TP rate for any category is 70% (*redundancy*) and highest FP rate is 20% (*word usage*). There are only two categories with either a TP rate less than 80% or a FP rate of more than 20%. On the other hand, the lowest TP is for the *subject* (70.6%) and highest FP is for *salutation* (17.5%) yet under the 20% mark. We ultimately aim to get all error rates in the 80-20 range of TP-FP.

6. CONCLUSION

We propose a system to grade email on content relevance, grammar and email etiquette using a mix of rule based and machine learning methods. We present a set of 36 quality metrics to evaluate email etiquette, their broad categorization and explanation. Our automated system provides human competitive performance on all evaluation parameters. The system provides scores on the three parameters and also detailed feedback on email etiquette. This feedback comprises the exact location of error, details of the error, and possible corrections.

In future, we plan to work on grading the finer aspects of email writing skills, such as flow of the email, its sentiment and how well the different parts of the email address the situation in the prompt. Also, this is a new interesting approach to process semi-structured text. We plan to use and benchmark the approach for other applications.

7. REFERENCES

[1] Yigal Attali and Jill Burstein. Automated essay scoring with e-rater v.2.0. *Journal of Technology, Learning, and Assessment*, 2006.

[2] Z. Chuang, W. Ming, L. C. Guang, X. Bo, and L. Zhi-qing. Resume parser: Semi-structured chinese document analysis. In *2009 WRI World Congress on Computer Science and Information Engineering*, volume 5, pages 12–16, March 2009.

[3] Maxine Eskenazi, , Maxine Eskenazi, and Scott Hansma. The fluency pronunciation trainer. In *Proc. Speech Technology in Language Learning 1998, Marholmen*, pages 77–80, 1998.

[4] Peter W. Foltz, Darrell Laham, and Thomas K Landauer. Automated essay scoring: Applications to educational technology. In Betty Collis and Ron Oliver, editors, *Proceedings of EdMedia + Innovate Learning 1999*, pages 939–944, Seattle, WA USA, 1999. Association for the Advancement of Computing in Education (AACE).

[5] National Center for O*NET Development. O*NET Resource Center. `https://www.onetcenter.org/overview.html`. Accessed: 2019-02-28.

[6] W. N. Francis and H. Kucera. Brown corpus manual. Technical report, Department of Linguistics, Brown University, Providence, Rhode Island, US, 1979.

[7] Horacio Franco, Harry Bratt, Romain Rossier, Venkata Gadde, Elizabeth Shriberg, Victor Abrash, and Kristin Precoda. Eduspeak(r): A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications. *Language Testing*, 27, 08 2010.

[8] Yong-Won Lee, Claudia Gentile, and Robert Kantor. Analytic Scoring of TOEFL CBT Essays: Scores from Humans and E-rater, 06 2008.

[9] Edward Loper and Steven Bird. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, pages 63–70, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[10] Angeliki Metallinou and Jian Cheng. Using deep neural networks to improve proficiency assessment for children english language learners. In *INTERSPEECH*, 2014.

[11] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[12] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November 1995.

[13] Michael Mohler, Razvan C. Bunescu, and Rada Mihalcea. Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In *ACL*, 2011.

[14] Lawrence M. Rudner and Tiankai Liang. Automated essay scoring using bayes' theorem. *Journal of Technology, Learning and Assessment, Vol. 1, No. 2.*, 2002.

[15] Shashank Srikant and Varun Aggarwal. A system to grade computer programming skills using machine learning. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1887–1896. ACM, 2014.

[16] Yla R. Tausczik and James W. Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54, 2010.

[17] Zhen Wang, Klaus Zechner, and Yu Sun. Monitoring the performance of human and automated scores for spoken responses. *Language Testing*, 35(1):101–120, 2018.

[18] Klaus Zechner, Derrick Higgins, Xiaoming Xi, and David M. Williamson. Automatic scoring of non-native spontaneous speech in tests of spoken english. *Speech Communication*, 51:883–895, 10 2009.