Eyes are the Windows to the Soul: Predicting the Rating of Text Quality Using Gaze Behaviour

★Sandeep Mathias, ★,♣,♢Diptesh Kanojia, ★Kevin Patel, ★Samarth Agrawal
♠Abhijit Mishra, ★Pushpak Bhattacharyya
★CSE Department, IIT Bombay
♣IITB-Monash Research Academy
♢Monash University, Australia
♠IBM Research, India
★,♣{sam,diptesh,kevin.patel,samartha,pb}@cse.iitb.ac.in
♠abhijimi@in.ibm.com

Abstract

Predicting a reader's rating of text quality is a challenging task that involves estimating different subjective aspects of the text, like structure, clarity, *etc.* Such subjective aspects are better handled using cognitive information. One such source of cognitive information is gaze behaviour. In this paper, we show that gaze behaviour does indeed help in effectively predicting the rating of text quality. To do this, we first model text quality as a function of three properties - organization, coherence and cohesion. Then, we demonstrate how capturing gaze behaviour helps in predicting each of these properties, and hence the overall quality, by reporting improvements obtained by adding gaze features to traditional textual features for score prediction. We also hypothesize that if a reader has fully understood the text, the corresponding gaze behaviour would give a better indication of the assigned rating, as opposed to partial understanding. Our experiments validate this hypothesis by showing greater agreement between the given rating and the predicted rating when the reader has a full understanding of the text.

1 Introduction

Automatically rating the quality of a text is an interesting challenge in NLP. It has been studied since Page's seminal work on automatic essay grading in the mid-1960s (Page, 1966). This is due to the dependence of quality on different aspects such as the overall structure of the text, clarity, *etc.* that are highly qualitative in nature, and whose scoring can vary from person to person (Person, 2013).

Scores for such qualitative aspects cannot be inferred solely from the text and would benefit from psycholinguistic information, such as gaze behaviour. Gaze based features have been used for co-reference resolution (Ross et al., 2016), sentiment analysis (Joshi et al., 2014) and translation annotation complexity estimation (Mishra et al., 2013). They could also be very useful for education applications, like evaluating readability (Mishra et al., 2017) and in automatic essay grading.

In this paper, we consider the following qualitative properties of text: Organization, Coherence and Cohesion. A text is **well-organized** if it begins with an introduction, has a body and ends with a conclusion. One of the other aspects of organization is the fact that it takes into account how the content of the text is split into paragraphs, with each paragraph denoting a single *idea*. If the text is too long, and not split into paragraphs, one could consider the text to be badly organized[1].

A text is **coherent** if it makes sense to a reader. A text is **cohesive** if it is well connected. Coherence and cohesion are two qualities that are closely related. A piece of text that is well-connected usually makes sense. Conversely, a piece of text that makes sense is usually well-connected. However, it is possible for texts to be coherent but lack cohesion. Table 1 provides some examples for texts that are coherent and cohesive, as well as those that lack one of those qualities.

There are different ways to model coherence and cohesion. Since coherence is a measure of how much sense the text makes, it is a semantic property of the text. It requires sentences within the text to be interpreted, by themselves, as well as with other sentences in the text (Van Dijk, 1980).

On the other hand, cohesion makes use of

[1]Refer supplementary material for example. We have placed it there due to space constraints.

Example	Comments
My favourite colour is blue. I like it because it is calming and it relaxes me. I often go outside in the summer and lie on the grass and look into the clear sky when I am stressed. For this reason, I'd have to say my favourite colour is blue.	Coherent and cohesive.
My favourite colour is blue. I'm calm and relaxed. In the summer I lie on the grass and look up.	Coherent but not cohesive. There is no link between the sentences. However, the text makes sense due to a lot of implicit clues (blue, favourite, relaxing, look up (and see the blue sky)).
My favourite colour is *blue*. *Blue* sports cars go **very fast**. Driving in **this way** is dangerous and can cause many *car crashes*. I had a *car accident* once and **broke my leg**. I was very sad because I had to miss a holiday in Europe because of **the injury**.	Cohesive but not coherent. The sentences are linked by words (that are in *italics* or in **bold**) between adjacent sentences. As we can see, every pair of adjacent sentences are connected by words / phrases, but the text does not make sense, since it first starts with blue, and describes missing a holiday due to injury.

Table 1: Examples of coherence and cohesion[2].

linguistic cues, such as references (demonstratives, pronouns, *etc.*), ellipsis (leaving out implicit words - Eg. Sam can type and I can [*type*] too), substitution (use of a word or phrase to replace something mentioned earlier - Eg. How's the croissant? I'd like to have **one** too.), conjunction (and, but, therefore, *etc.*), cohesive nouns (problem, issue, investment, *etc.*) and lexis (linking different pieces of text by synonyms, hyponyms, lexical chains, *etc.*) (Halliday and Hasan, 1976).

Using these properties, we model the overall text quality rating. We make use of a Likert scale (Likert, 1932) with a range of 1 to 4, for measuring each of these properties; the higher the score, the better is the text in terms of that property. We model the text quality rating on a scale of 1 to 10, using the three scores as input. In other words,

$$Quality(T) = Org(T) + Chr(T) + Chs(T) - 2,$$

where $Quality(T)$ is the text quality rating of the text T. $Org(T)$, $Chr(T)$, and $Chs(T)$ correspond to the **Org**anization, **C**o**h**e**r**ence, and **C**o**h**e**s**ion scores respectively, for the text T, that are given by a reader. We subtract 2 to scale the scores from a range of 3 - 12, to a range of 1 - 10 for quality.

Texts with poor organization and/or cohesion can force readers to regress *i.e.* go to previous sentences or paragraphs. Texts with poor coherence may lead readers to fixate more on different portions of text to understand them. In other words, such gaze behaviour indirectly captures the effort needed by human readers to comprehend the text (Just and Carpenter, 1980), which, in turn, may influence the ratings given by them. Hence, these properties seem to be a good indicators for overall quality of texts.

In this paper, we address the following question: ***Can information obtained from gaze behaviour help predict reader's rating of quality of text by estimating text's organization, coherence, and cohesion?*** Our work answers that question in the affirmative. We found that using gaze features does contribute in improving the prediction of qualitative ratings of text by users.

Our work has the following contributions. Firstly, we propose **a novel way to predict readers' rating of text** by recording their eye movements as they read the texts. Secondly, we show that **if a reader has understood the text completely, their gaze behaviour is more reliable**. Thirdly, **we also release our dataset**[3] to help in further research in using gaze features in other tasks involving predicting the quality of texts.

In this paper, we use the following terms related to eye tracking. The **interest area (IA)** is an area of the screen that is under interest. We mainly look at words as interest areas. A **fixation** takes place when the gaze is focused on a point of the screen. A **saccade** is the movement of gaze between two fixations. A **regression** is a special type of saccade in which the reader refers back to something that they had read earlier.

The rest of the paper is organized as follows. Section 2 describes the motivation behind our work. Section 3 describes related work in this field. Section 4 describes the different features that we used. Sections 5 and 6 describes our experiments and results. Section 6 also contains analysis of our experiments. Section 7 concludes our paper and mentions future work.

[2]We took the examples from this site for explaining coherence and cohesion: http://gordonscruton.blogspot.in/2011/08/what-is-cohesion-coherence-cambridge.html

[3]The dataset can be downloaded from http://www.cfilt.iitb.ac.in/cognitive-nlp/

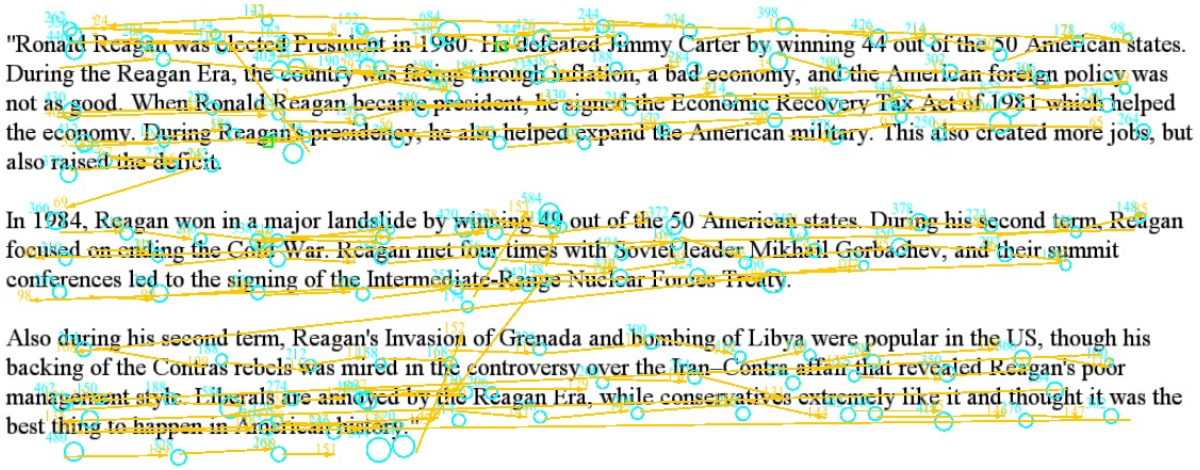


Figure 1: Sample text showing fixations, saccades and regressions. This text was given scores of 4, 4, and 3 for organization, coherence and cohesion. The circles denote fixations, and the lines are saccades. Radius of the circles denote the duration of the fixation (in milliseconds), which is centred at the centre of the circle. This is the output from SR Research Data Viewer software.

2 Motivation

Reader's perception of text quality is subjective and varies from person to person. Using cognitive information from the reader can help in predicting the score he / she will assign to the text. A well-written text would not have people fixate too long on certain words, or regress a lot to understand, while a badly written text would do so.

Figure 1 shows the gaze behaviour for a sample text. The circles denote fixations, and the arrows denote saccades. If we capture the gaze behaviour, as well as see how well the reader has understood the text, we believe that we can get a clearer picture of the quality rating of the text.

One of the major concerns is *How are we going to get the gaze data?* This is because capability to gather eye-tracking data is not available to the masses. However, top mobile device manufacturers, like Samsung, have started integrating basic eye-tracking software into their smartphones (Samsung Smart Scroll) that are able to detect where the eye is fixated, and can be used in applications like scrolling through a web page. Start-ups, like Cogisen[4], have started using gaze features in their applications, such as using gaze information to improve input to image processing systems. Recently, SR Research has come up with a portable eye-tracking system[5].

[4] www.cogisen.com

[5] https://www.sr-research.com/products/eyelink-portable-duo/

3 Related Work

A number of studies have been done showing how eye tracking can model aspects of text. Word length has been shown to be positively correlated with fixation count (Rayner, 1998) and fixation duration (Henderson and Ferreira, 1993). Word predictability (i.e. how well the reader can predict the next word in a sentence) was also studied by Rayner (1998), where he found that unpredictable words are less likely to be skipped than predictable words.

Shermis and Burstein (2013) gives a brief overview of how text-based features are used in multiple aspects of essay grading, including grammatical error detection, sentiment analysis, short-answer scoring, *etc.* Their work also describes a number of current essay grading systems that are available in the market like *E-rater*® (Attali and Burstein, 2004). In recent years, there has been a lot of work done on evaluating the holistic scores of essays, using deep learning techniques (Alikaniotis et al., 2016; Taghipour and Ng, 2016; Dong and Zhang, 2016).

There has been little work done to model text organization, such as Persing et al. (2010) (using machine learning) and Taghipour (2017) (using neural networks). However, there has been a lot of work done to model coherence and cohesion, using methods like lexical chains (Somasundaran et al., 2014), an entity grid (Barzilay and Lapata, 2005), *etc.* An interesting piece of work to model coherence was done by Soricut and Marcu (2006)

where they used a machine translation-based approach to model coherence. Zesch et al. (2015) use topical overlap to model coherence for essay grading. Discourse connectors are used as a heuristic to model cohesion by Zesch et al. (2015) and Persing and Ng (2015). Our work is novel because it makes use of gaze behaviour to model and predict coherence and cohesion in text.

In recent years, there has been some work in using eye-tracking to evaluate certain aspects of the text, like readability (Gonzalez-Garduño and Søgaard, 2017; Mishra et al., 2017), grammaticality (Klerke et al., 2015), *etc.*. Our work uses eye-tracking to predict the score given by a reader to a complete piece of text (rather than just a sentence as done by Klerke et al. (2015)) and show that the scoring is more reliable if the reader has understood the text.

4 Features

In order to predict the scores of the different properties of the text, we use the following text and gaze features.

4.1 Text-based Features

We use a set of text-based features to come up with a baseline system to predict the scores for different properties.

The first set of features that we use are **length and count-based features**, such as word length, word count, sentence length, count of transition phrases[6] *etc.* (Persing and Ng, 2015; Zesch et al., 2015).

The next set of features that we use are **complexity features**, namely the degree of polysemy, coreference distance, and the Flesch Reading Ease Score (FRES) (Flesch, 1948). These features help in normalizing the gaze features for text complexity. These features were extracted using Stanford CoreNLP (Manning et al., 2014), and MorphAdorner (Burns, 2013).

The third set of features that we use are **stylistic features** such as the ratios of the number of adjectives, nouns, prepositions, and verbs to the number of words in the text. These features are used to model the distributions of PoS tags in good and bad texts. These were extracted using NLTK[7] (Loper and Bird, 2002).

The fourth set of features that we use are **word embedding features**. We use the average of word vectors of each word in the essay, using Google News word vectors (Mikolov et al., 2013). The word embeddings are **300 dimensions**. We also calculate the mean and maximum similarities between the word vectors of the content words in adjacent sentences of the text, using GloVe word embeddings[8] (Pennington et al., 2014).

The fifth set of features that we use are **language modeling features**. We use the count of words that are absent in Google News word vectors and misspelled words using the PyEnchant[9] library. In order to check the grammaticality of the text, we construct a 5-gram language model, using the Brown Corpus (Francis and Kucera, 1979).

The sixth set of features are **sequence features**. These features are particularly useful in modeling organization (sentence and paragraph sequence similarity) (Persing et al., 2010), coherence and cohesion (PoS and lemma similarity). Pitler et al. (2010) showed that cosine similarity of adjacent sentences as one of the best predictors of linguistic quality. Hence, we also create vectors for the PoS tags and lemmas for each sentence in the text. The dimension of the vector is the number of distinct PoS tags / lemmas.

The last set of features that we look at are **entity grid features**. We define entities as the nouns in the document, and do coreference resolution to resolve pronouns. We then construct an entity grid (Barzilay and Lapata, 2005) - a 1 or 0 grid that checks whether an entity is present or not in a given sentence. We take into account sequences of entities across sentences that possess ***at least*** one 1, that are either bigrams, trigrams or 4-grams. A sequence with multiple 1s denote entities that are close to each other, while sequences with a solitary 1 denote that an entity is just mentioned once and we do not come across it again for a number of sentences.

4.2 Gaze-based Features

The gaze-based features are dependent on the gaze behaviour of the participant with respect to interest areas.

[6] https://writing.wisc.edu/Handbook/Transitions.html

[7] http://www.nltk.org/

[8] We found that using GloVe here and Google News for the mean word vectors worked best.

[9] https://pypi.python.org/pypi/pyenchant/

Fixation Features

The **First Fixation Duration** (FFD) shows the time the reader fixates on a word when he / she first encounters it. An increased FFD intuitively could mean that the word is more complex and the reader spends more time in understanding the word (Mishra et al., 2016).

The **Second Fixation Duration** (SFD) is the duration in which the reader fixates on a particular interest area the second time. This happens during a regression, when a reader is trying to link the word he / she just read with an earlier word.

The **Last Fixation Duration** (LFD) is the duration in which the reader fixates on a particular interest area the final time. At this point, we believe that the interest area has been processed.

The **Dwell Time** (DT) is the total time the reader fixates on a particular interest area. Like first fixation, this also measures the complexity of the word, not just by itself, but also with regard to the entire text (since it takes into account fixations when the word was regressed, *etc.*)

The **Fixation Count** (FC) is the number of fixations on a particular interest area. A larger fixation count could mean that the reader frequently goes back to read that particular interest area.

Regression Features

IsRegression (IR) is the number of interest areas where a regression happened before reading ahead and **IsRegressionFull** (IRF) is the number of interest areas where a regression happened. The **Regression Count** (RC) is the total number of regressions. The **Regression Time** (RT) is the duration of the regressions from an interest area. These regression features could help in modeling semantic links for coherence and cohesion.

Interest Area Features

The **Skip Count** (SC) is the number of interest areas that have been skipped. The **Run Count** (RC) is the number of interest areas that have at least one fixation. A larger run count means that more interest areas were fixated on. Badly written texts would have higher run counts (and lower skip counts), as well as fixation counts, because the reader will fixate on these texts for a longer time to understand them.

5 Experiment Details

In this section, we describe our experimental setup, creation of the dataset, evaluation metric, classifier details, *etc.*

5.1 Ordinal Classification vs. Regression

For each of the properties - organization, coherence and cohesion, we make use of a Likert scale, with scores of 1 to 4. Details of the scores are given in Table 2. For scoring the quality, we use the formula described in the Introduction. Since we used a Likert scale, we make use of ordinal classification, rather than regression. This is because each of the grades is a discrete value that can be represented as an ordinal class (where $1 < 2 < 3 < 4$), as compared to a continuous real number.

5.2 Evaluation Metric

For the predictions of our experiments, we use Cohen's Kappa with quadratic weights - quadratic weighted Kappa (QWK) (Cohen, 1968) because of the following reasons. Firstly, unlike accuracy and F-Score, Cohen's Kappa takes into account whether or not agreements happen by chance. Secondly, weights (either linear or quadratic) take into account distance between the given score and the expected score, unlike accuracy and F-score where mismatches (either 1 vs. 4, or 1 vs.2) are penalized the same. Quadratic weights reward matches and penalize mismatches more than linear weights.

To measure the Inter-Annotator Agreement of our raters, we make use of Gwet's second-order agreement coefficient (Gwet's AC2) as it can handle ordinal classes, weights, missing values, and multiple raters rating the same document (Gwet, 2014).

5.3 Creation of the Dataset

In this subsection, we describe how we created our dataset. We describe the way we made the texts, the way they were annotated and the inter-annotator agreements for the different properties.

Details of Texts

To the best of our knowledge there isn't a publicly available dataset with gaze features for textual quality. Hence, we decided to create our own. Our dataset consists of a diverse set of **30 texts**, from Simple English Wikipedia (**10 articles**), English Wikipedia (**8 articles**), and online news articles (**12 articles**)[10]. We did not wish to overburden the readers, so we kept the size of texts to

[10]The sources for the articles were https://simple.wikipedia.org, https://en.wikipedia.org, and https://newsela.com

Property	Grade	Guidelines
Organization	1	**Bad**. There is no organization in the text.
	2	**OK**. There is little / no link between the paragraphs, but they each describe an idea.
	3	**Good**. Some paragraphs may be missing, but there is an overall link between them.
	4	**Very Good**. All the paragraphs follow a flow from the Introduction to Conclusion.
Coherence	1	**Bad**. The sentences do not make sense.
	2	**OK**. Groups of sentences may make sense together, but the text still may not make sense.
	3	**Good**. Most of the sentences make sense. The text, overall, makes sense.
	4	**Very Good**. The sentences and overall text make sense.
Cohesion	1	**Bad**. There is little / no link between any 2 adjacent sentences in the same paragraph.
	2	**OK**. There is little / no link between adjacent paragraphs. However, each paragraph is cohesive
	3	**Good**. All the sentences in a paragraph are linked to each other and contribute in understanding the paragraph.
	4	**Very Good**. The text is well connected. All the sentences are linked to each other and help in understanding the text.

Table 2: Annotation guidelines for different properties of text.

approximately **200 words** each. The original articles ranged from a couple hundred words (Simple English Wikipedia) to over a thousand words (English Wikipedia). We first summarized the longer articles manually. Then, for the many articles over 200 words, we removed a few of the paragraphs and sentences. In this way, despite all the texts being published, we were able to introduce some poor quality texts into our dataset. The articles were sampled from a variety of genres, such as History, Science, Law, Entertainment, Education, Sports, *etc.*

Details of Annotators

The dataset was annotated by **20 annotators** in the **age** group of **20-25**. Out of the 20 annotators, the distribution was 9 high school graduates (current college students), 8 college graduates, and 3 annotators with a post-graduate degree.

In order to check the **eyesight** of the annotators, we had each annotator look at different parts of the screen. While they did that, we recorded how their fixations were being detected. Only if their fixations to particular parts of the screen tallied with our requests, would we let them participate in annotation.

All the participants in the experiment were **fluent speakers of English**. A few of them scored over 160 in GRE Verbal test and/or over 110 in TOEFL. Irrespective of their appearance in such exams, each annotator was made to take an English test before doing the experiments. The participants had to read a couple of passages, answer comprehension questions and score them for organization, coherence and cohesion (as either good / medium / bad). In case they either got both comprehension questions wrong, or labeled a good passage bad (or vice versa), they failed the test[11].

[11]25 annotators applied, but we chose only 20. 2 of the rejected annotators failed the test, while the other 3 had bad eyesight.

Property	Full	Overall
Organization	0.610	0.519
Coherence	0.688	0.633
Cohesion	0.675	0.614

Table 3: Inter-Annotator Agreements (Gwet's AC2) for each of the properties.

In order to help the annotators, they were given 5 **sample texts** to differentiate between good and bad organization, coherence and cohesion. Table 1 has some of those texts[12].

Inter-Annotator Agreement

Each of the properties were scored in the range of 1 to 4. In addition, we also evaluated the participant's understanding of the text by asking them a couple of questions on the text. Table 3 gives the inter-annotator agreement for each of the 3 properties that they rated. The column **Full** shows the agreement only if the participant answered both the questions correct. The **Overall** column shows the agreement irrespective of the participant's comprehension of the text.

5.4 System Details

We conducted the experiment by following standard norms in eye-movement research (Holmqvist et al., 2011). The display screen is kept **about 2 feet** from the reader, and the camera is placed midway between the reader and the screen. The reader is seated and the position of his head is fixed using a chin rest.

Before the text is displayed, we calibrate the camera by having the participant fixate on **13**

[12]The texts for good and bad organization are too long to provide in this paper. They will be uploaded in supplementary material.

points on the screen and validate the calibration so that the camera is able to predict the location of the eye on the screen accurately. After calibration and validation, the text is displayed on the screen in **Times New Roman** typeface with **font size 23**. The reader reads the text and while that happens, we record the reader's eye movements. The readers were allowed to take **as much time as they needed** to finish the text. Once the reader has finished, the reader moves to the next screen.

The next two screens each have a question that is based on the passage. These questions are used to verify that the reader did not just skim through the passage, but understood it as well. The questions were multiple choice, with 4 options[13]. The questions test literal comprehension (where the reader has to recall something they read), and interpretive comprehension (where the reader has to infer the answer from the text they read). After this, the reader scores the texts for organization, coherence and cohesion. The participants then take a short break (about 30 seconds to a couple of minutes) before proceeding with the next text. This is done to prevent reading fatigue over a period of time. After each break, we recalibrate the camera and validate the calibration again.

For obtaining gaze features from a participant, we collect gaze movement patterns using an SR Research Eye Link 1000 eye-tracker (monocular stabilized head mode, sampling rate 500Hz). It is able to collect all the gaze details that we require for our experiments. Reports are generated for keyboard events (message report) and gaze behaviour (interest area report) using SR Research Data Viewer software.

5.5 Classification Details

We also process the articles for obtaining the text features as described in Section 4. Given that we want to show the utility of gaze features, we ran each of the following classifiers with 3 feature sets - only text, only gaze, and all features.

We split the data into a training - test split of sizes **70%** and **30%**. We used a Feed Forward Neural Network with **1 hidden layer** containing **100 neurons** (Bebis and Georgiopoulos, 1994)[14]. The size of the input vector was **361 features**. Out of these, there were **49 text features**, plus **300 dimension word embeddings features**, **11 gaze features**, and **1 class label**. The data was split using stratified sampling, to ensure that there is a similar distribution of classes in each of the training and test splits. The Feed Forward Neural Network was implemented using TensorFlow (Abadi et al., 2015) in Python. We ran the neural network over **10,000 epochs**, with a learning rate of **0.001** in **10 batches**. The loss function that we used was the **mean square error**.

In order to see how much the participant's understanding of the text would reflect on their scoring, we also looked at the data based on how the participant scored in the comprehension questions after they read the article. We split the articles into 2 subsets here - $Full$, denoting that the participant answered both the questions correctly, and $Partial$, denoting that they were able to answer only one of the questions correctly. The readers showed $Full$ understanding in **269 instances** and $Partial$ understanding in **261 instances**. We used the same setup here (same training - test split, stratified sampling, and feed forward neural network). We omit the remaining **70 instances** where the participant got none of the questions correct, as the participant could have scored the texts completely randomly.

6 Results and Analysis

Table 4 shows the results of our experiments using the feed forward neural network classifier. The first column is the property being evaluated. The next 3 columns denote the results for the Text, Gaze and Text+Gaze feature sets.

Property	Text	Gaze	Text+Gaze
Organization	0.237	0.394	**0.563**
Coherence	0.261	0.285	**0.550**
Cohesion	0.120	0.229	**0.451**
Quality	0.230	0.304	**0.552**

Table 4: QWK scores for the three feature sets on different properties.

The QWK scores are the predictions which we obtain with respect to the scores of all the 30 documents, scored by all 20 raters. Textual features when augmented with gaze based features show significant improvement for all the properties.

[13] **Example Passage Text**: The text in Figure 1
Question: "How many states did Ronald Reagan win in both his Presidential campaigns?"
Correct Answer: "93" (44+49)

[14] We also used other classifiers, like Naive Bayes, Logistic Regression and Random Forest. However, the neural network outperformed them.

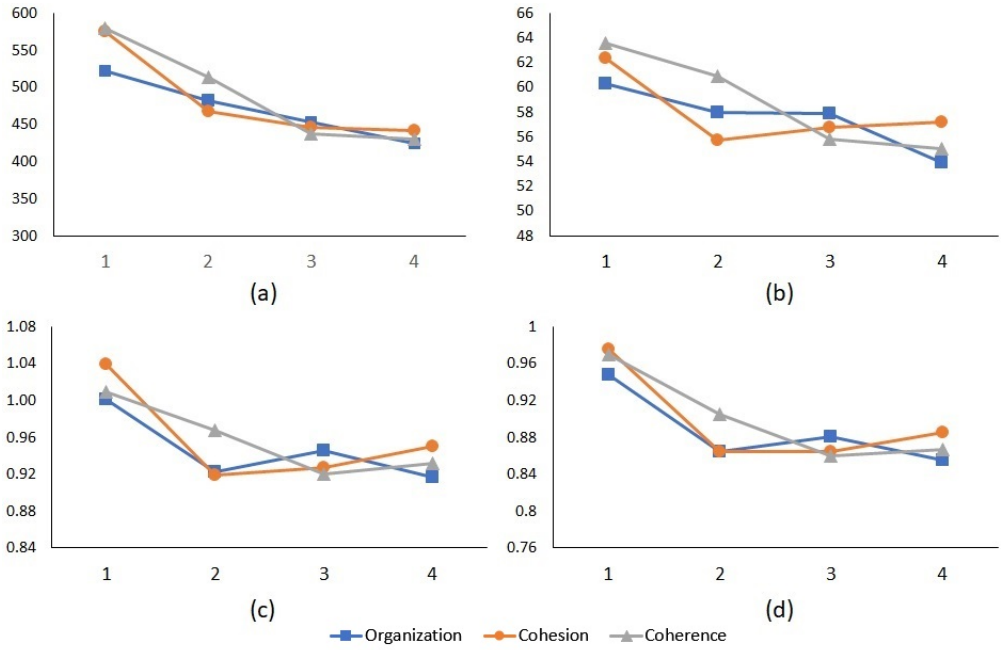


Figure 2: Relation between some of the different gaze features and the score. The gaze features are (a) RD, (b) SFD, (c) FC and (d) RC. For figures (a) and (b), the units on the y-axis are milliseconds. For figures (c) and (d) the numbers are a ratio to the number of interest areas in the text. The x-axis in all 4 graphs is the score given by the annotators.

We check the statistical significance of improvement of adding gaze based features for the results in Table 4. To test our hypothesis - that adding gaze features make a statistically significant improvement - we run the t-test. Our null hypothesis: Gaze based features do not help in prediction, any more than text features themselves, and whatever improvements happen when gaze based features are added to the textual features, are not statistically significant. We choose a significance level of $p < 0.001$. For all the improvements, we found them to be statistically significant above this α level, rejecting our null hypothesis.

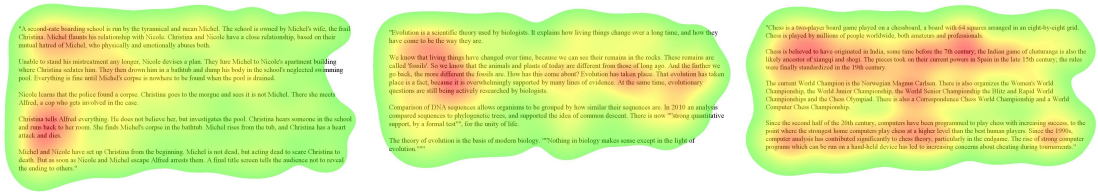
We also evaluate how the participant's understanding of the text affects the way they score the text. Table 5 shows the results of our experiments taking the reader's comprehension into account. The first column is the property being evaluated. The second column is the level of comprehension - $Full$ for the passages where the participant answered both the questions correctly, and $Partial$ for the passages where the participant answered one question correctly. The next 3 columns show the results using the Text feature set, the Gaze feature set, and both (Text+Gaze) feature sets. From this table, we see that wherever the gaze features are used, there is far greater agreement for those with $Full$ understanding as compared to $Partial$ understanding.

Property	Comp.	Text	Gaze	Text+Gaze
Organization	Full	**0.319**	**0.319**	**0.563**
	Partial	0.115	0.179	0.283
Coherence	Full	0.255	**0.385**	**0.601**
	Partial	**0.365**	0.343	0.446
Cohesion	Full	**0.313**	**0.519**	**0.638**
	Partial	0.161	0.155	0.230
Quality	Full	**0.216**	**0.624**	**0.645**
	Partial	0.161	0.476	0.581

Table 5: QWK scores for the three feature sets on different properties categorized on the basis of reader comprehension.

Figure 2 shows a clear relationship between some of the gaze features and the scores given by readers for the properties - organization, cohesion and coherence. In all the charts, we see that texts with the lowest scores have the longest durations (regression / fixation) as well as counts (of fixations and interest areas fixated).

Figure 3 shows the fixation heat maps for 3 texts whose quality scores were good (10), medium (6) and bad (3), read by the same participant. From these heat maps, we see that the text rated good has highly dense fixations for only a part of the text,



(a) Good (rated 10) (b) Medium (rated 6) (c) Bad (rated 3)

Figure 3: Fixation heatmap examples for one of the participants from SR Research Data Viewer software.

as compared to the medium and bad texts. This shows that badly written texts force the readers to fixate a lot more than well-written texts.

6.1 Ablation Tests

In order to see which of the gaze feature sets is important, we run a set of ablation tests. We ablate the fixations, regressions and interest area feature sets one at a time. We also ablated each of the **individual gaze features**.

Property	Fixation	Regression	Interest Areas
Organization	-0.102	-0.017	**-0.103**
Coherence	-0.049	-0.077	**-0.088**
Cohesion	-0.015	**-0.040**	0.037
Quality	0.002	0.016	**-0.056**

Table 6: **Difference** in QWK scores when ablating three gaze behaviour feature sets for different properties.

Table 6 gives the result of our ablation tests on the three feature sets - fixation, regression and interest area feature sets. The first column is the property that we are measuring. The next 3 columns denote the **difference** between the predicted QWK that we got from ablating the fixation, regression and interest area feature sets. We found that the Interest Area feature set was the most important, followed by fixation and regression.

Among the individual features, **Run Count** (RC) was found to be the most important for organization and quality. **First Fixation Duration** (FFD) was the most important feature for coherence, and **IsRegressionFull** (IRF) was the most important feature for cohesion. We believe that this is because the number of interest areas that are fixated on at least once and the number of interest areas that are skipped play an important role in determining how much of the text was read and how much was skipped. However, for cohesion, regression features are the most important, because they show a link between the cohesive clues (like lexis, references, *etc.*) in adjacent sentences.

7 Conclusion and Future Work

We presented a novel approach to predict reader's rating of texts. The approach estimates the overall quality on the basis of three properties - organization, coherence and cohesion. Although well defined, predicting the score of these properties for a text is quite challenging. It has been established that cognitive information such as gaze behaviour can help in such subjective tasks (Mishra et al., 2013, 2016). We hypothesized that gaze behavior will assist in predicting the scores of text quality. To evaluate this hypothesis, we collected gaze behaviour data and evaluated the predictions using only the text-based features. When we took gaze behaviour into account, we were able to significantly improve our predictions of organization, coherence, cohesion and quality. We found out that, in all cases, there was an improvement in the agreement scores when the participant who rated the text showed full understanding, as compared to partial understanding, using only the Gaze features and the Text+Gaze features. This indicated that gaze behaviour is more reliable when the reader has understood the text.

To the best of our knowledge, our work is pioneering in using gaze information for predicting text quality rating. In future, we plan to use use approaches, like multi-task learning (Mishra et al., 2018), in estimating gaze features and using those estimated features for text quality prediction.

Acknowledgements

We'd like to thank all the anonymous reviewers for their constructive feedback in helping us improve our paper. We'd also like to thank Anoop Kunchukuttan, a research scholar from the Centre for Indian Language Technology, IIT Bombay for his valuable input.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org. https://www.tensorflow.org/.

Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. 2016. Automatic text scoring using neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 715–725.

Yigal Attali and Jill Burstein. 2004. Automated essay scoring with e-rater® v. 2.0. *ETS Research Report Series* 2004(2).

Regina Barzilay and Mirella Lapata. 2005. Modeling local coherence: An entity-based approach. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. Association for Computational Linguistics, Ann Arbor, Michigan, pages 141–148. https://doi.org/10.3115/1219840.1219858.

George Bebis and Michael Georgiopoulos. 1994. Feed-forward neural networks. *IEEE Potentials* 13(4):27–31.

Philip R Burns. 2013. Morphadorner v2: A java library for the morphological adornment of english language texts. *Northwestern University, Evanston, IL* .

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin* 70(4):213.

Fei Dong and Yue Zhang. 2016. Automatic features for essay scoring – an empirical study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 1072–1077.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology* 32(3):221.

W Nelson Francis and Henry Kucera. 1979. The brown corpus: A standard corpus of present-day edited american english. *Providence, RI: Department of Linguistics, Brown University [producer and distributor]* .

Ana Valeria Gonzalez-Garduño and Anders Søgaard. 2017. Using gaze to predict text readability. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, Copenhagen, Denmark, pages 438–443. http://www.aclweb.org/anthology/W17-5050.

Kilem L Gwet. 2014. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Michael Alexander Kirkwood Halliday and Ruqaiya Hasan. 1976. *Cohesion in english*. Longman Group Ltd.

John M Henderson and Fernanda Ferreira. 1993. Eye movement control during reading: Fixation measures reflect foveal but not parafoveal processing difficulty. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* 47(2):201.

Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost Van de Weijer. 2011. *Eye tracking: A comprehensive guide to methods and measures*. OUP Oxford.

Aditya Joshi, Abhijit Mishra, Nivvedan Senthamilselvan, and Pushpak Bhattacharyya. 2014. Measuring sentiment annotation complexity of text. In *ACL (2)*. pages 36–41.

Marcel A Just and Patricia A Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological review* 87(4):329.

Sigrid Klerke, Héctor Martínez Alonso, and Anders Søgaard. 2015. Looking hard: Eye tracking for detecting grammaticality of automatically compressed sentences. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*. Linköping University Electronic Press, Sweden, Vilnius, Lithuania, pages 97–105. http://www.aclweb.org/anthology/W15-1814.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology* .

Edward Loper and Steven Bird. 2002. 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*. Association for Computational Linguistics, pages 63–70.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*. pages 55–60. http://www.aclweb.org/anthology/P/P14/P14-5010.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pages 3111–3119.

Abhijit Mishra, Pushpak Bhattacharyya, and Michael Carl. 2013. Automatically predicting sentence translation difficulty. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Sofia, Bulgaria, pages 346–351.

Abhijit Mishra, Diptesh Kanojia, and Pushpak Bhattacharyya. 2016. Predicting readers' sarcasm understandability by modeling gaze behavior. In *AAAI*. pages 3747–3753.

Abhijit Mishra, Diptesh Kanojia, Seema Nagar, Kuntal Dey, and Pushpak Bhattacharyya. 2017. Scanpath complexity: Modeling reading effort using gaze information. In *AAAI*. pages 4429–4436.

Abhijit Mishra, Srikanth Tamilselvam, Riddhiman Dasgupta, Seema Nagar, and Kuntal Dey. 2018. Cognition-cognizant sentiment analysis with multitask subjectivity summarization based on annotators gaze behavior .

Ellis B Page. 1966. The imminence of... grading essays by computer. *The Phi Delta Kappan* 47(5):238–243.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pages 1532–1543.

Isaac Persing, Alan Davis, and Vincent Ng. 2010. Modeling organization in student essays. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Cambridge, MA, pages 229–239.

Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Beijing, China, pages 543–552. http://www.aclweb.org/anthology/P15-1053.

Robert Person. 2013. Blind truth: An examination of grading bias.

Emily Pitler, Annie Louis, and Ani Nenkova. 2010. Automatic evaluation of linguistic quality in multidocument summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Uppsala, Sweden, pages 544–554. http://www.aclweb.org/anthology/P10-1056.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin* 124(3):372.

Joe Cheri Ross, Abhijit Mishra, and Pushpak Bhattacharyya. 2016. Leveraging annotators gaze behaviour for coreference resolution. In *Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning*. pages 22–26.

Mark D Shermis and Jill Burstein. 2013. *Handbook of automated essay evaluation: Current applications and new directions*. Routledge.

Swapna Somasundaran, Jill Burstein, and Martin Chodorow. 2014. Lexical chaining for measuring discourse coherence quality in test-taker essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin City University and Association for Computational Linguistics, Dublin, Ireland, pages 950–961.

Radu Soricut and Daniel Marcu. 2006. Discourse generation using utility-trained coherence models. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*. Association for Computational Linguistics, Sydney, Australia, pages 803–810. http://www.aclweb.org/anthology/P/P06/P06-2103.

Kaveh Taghipour. 2017. *Robust Trait-Specific Essay Scoring Using Neural Networks and Density Estimators*. Ph.D. thesis.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 1882–1891.

Teun Adrianus Van Dijk. 1980. Text and context explorations in the semantics and pragmatics of discourse .

Torsten Zesch, Michael Wojatzki, and Dirk Scholten-Akoun. 2015. Task-independent features for automated essay grading. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, Denver, Colorado, pages 224–232. http://www.aclweb.org/anthology/W15-0626.