Predicting the Readability of Short Web Summaries

Tapas Kanungo
kanungo@yahoo-inc.com
Yahoo! Labs
2821 Mission College Blvd.
Santa Clara, CA 95054

David Orr
dmorr@yahoo-inc.com
Yahoo! Labs
2821 Mission College Blvd.
Santa Clara, CA 95054

ABSTRACT

Readability is a crucial presentation attribute that web summarization algorithms consider while generating a query-baised web summary. Readability quality also forms an important component in real-time monitoring of commercial search-engine results since readability of web summaries impacts clickthrough behavior, as shown in recent studies, and thus impacts user satisfaction and advertising revenue.

The standard approach to computing the readability is to first collect a corpus of random queries and their corresponding search result summaries, and then each summary is then judged by a human for its readabilty quality. An average readability score is then reported. This process is time consuming and expensive. Besides, the manual evaluation process can not be used in the real-time summary generation process. In this paper we propose a machine learning approach to the problem. We use the corpus as described above and extract summary features that we think may characterize readability. We then estimate a model (gradient boosted decision tree) that predicts human judgments given the features. This model can then be used in real time to estimate the readability of new (unseen) web search summaries and also be used in the summary generation process.

We present results on approximately 5000 editorial judgments collected over the course of a year and show examples where the model predicts the quality well and where it disagrees with human judgments. We compare the results of the model to previous models of readability, most notably Collins-Thompson-Callan, Fog and Flesch-Kincaid, and see that our model shows substantially better correlation with editorial judgments as measured by Pearson's correlation coefficient. The learning algorithm also provides us with the relative importance of the features used.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

WSDM '09, February 9-12, 2009, Barcelona, Spain.

General Terms

Algorithms, Experimentation, Theory

Keywords

Readability, summarization, gradient boosted decision trees, realtime

1. INTRODUCTION

Web search is now a critical part of our everyday life. However, with the rapid growth of information on the web, finding a relevant web page is becoming more challenging. One of the standard approaches that the commercial search engines use to reduce a user's search time is to display a *summary* (see Figure 1) of each web page in the result set. Interestingly, a recent study [6] has shown that the readability of a summary has a direct impact on the click behavior on the search result page. A less readable summary is less likely to generate a click than a more readable summary. Thus it is important for search engines to generate nice, human readable, summaries.

Most search engines monitor various metrics that reflect some aspect of relevance. In fact, besides monitoring the relevance of their own engine, they monitor their competitors too. Some of the click-based relevance metrics that are monitored are click-through rate at various result positions, rate at which users return back to the result page immediately, etc. The primary task of a search result summary is to convey to the user the relevance (or lack thereof) of the result.

However, there are other dimensions of quality of a summary. The summary should convey information to the user in the most easily accessible way possible, and one major factor in accessibility to a user is readability. It is therefore important for search engines to monitor the readability of summaries on a regular basis. Such readability metrics would allow us to make sure that the result summary quality is not degrading, and that the readability is comparable to competitors. In addition, the readability monitor can serve as a way to detect sudden system changes that have a negative impact on the readability.

The current methodology of evaluating summary quality is, however, a very slow process. Queries are first sampled randomly from a weekly query log. The queries are then issued to various search engines and top-k (usually k is 10) results collected. Human judges rate the readability of the summaries (along with other quality judgments such as relevance) and various metrics are reported. Since this process is

Figure 1: A search result page summary. The summary is composed of a title, an abstract and a URL. The abstract itself can be composed of one or more "snippets." The snippets are either complete sentences or fragments of sentences. If a snippet is a fragment, ellipses are used to represent the missing chunks.

labor intensive, it is conducted once a quarter at best. (The process is quite similar to the TREC evaluations.) However, having an estimate of the human judgment on a daily basis would be very useful, as discussed earlier.

We propose a machine learning approach to modeling and monitoring the readability of search result summaries. Our approach is as follows. We collect a corpus of search result summaries and corresponding human judgments as described above. The judgments are from 1-5 where 1 is poor and 5 is good. Then we extract various features from the summaries and model the judgments as function of the features. The modeling itself is done using stochastic gradient boosted decision trees (GBDT) [8, 9]. This model is then used for rating a collection of new search result summaries.

The remainder of this paper is organized as follows. In Section 2, we discuss papers that have addressed similar issues and then identify why they are not appropriate for evaluating readability on web search summaries. Then, in Section 3, we describe our gradient boosted decision tree modeling framework and the readability features that we use. In Section 4 we describe our experimental approach, which includes the training and testing corpus, human judgment tools and guidelines, as well as choice of parameter values used in the GBDT algorithm. Then in Section 5 we present our results and discuss in what situations the readability predictions produced by our algorithm do not agree with human judges, and consider possible approaches to improving the model. We also discuss how this work is used for real time monitoring.

2. RELATED WORK

In a recent paper, Clarke *et al.* [6] studied the relationship of "click inversions" – rank locations where a lower ranked page gets more clicks than the previous one – and various attributes of search result page summaries. One of the conclusions of their work was that the readability of the abstract was statistically correlated with the clickthrough rates. However, the model of readability itself was heuristic and not tested against human readability judgments.

Numerous researchers have conducted user studies to understand aspects of readability. Aula [3] conducted user studies and found that putting each snippet on a separate line reduces the time humans take to process an abstract. Radev and Fan [23] found that readability of summaries decreased with the amount of compression. In addition they found that the processing time by humans increased when the compression ratio increased, most likely due to decrease in readability. Obendorf and Weinreich [22] found that the current visualization of hyperlinks can reduce the human readability of text considerably. Kickmeier and Albert [15] found that bolding of keywords in the summary is important for "scannability;" that is, bolding helps in getting a particular result noticed. Finally, Rose *et al.* [27] found that text choppiness and truncation affected the readability for human subjects.

All the user studies mentioned above, in general, give tasks to human subjects and measure some sort of metric (e.g. time to finish a task) and arrive at conclusions. Their methodology does not come up with a model of readability, and thus can not be used to predict human readability judgments in realtime.

FOG [10], Flesch, Flesch-Kincaid [16] and SMOG [20] are computational models for predicting readability of texts. Automated essay scoring systems use some of these methods [5]. Most of these methods assume long, well written text, and extract features like average number of characters per word and average syllables per word and predict the score using a linear model.

Similarly, Si and Callan [29] built a classifier using a linear combination of models based on language models and surface features. They too assumed complete sentences and not summaries which can contain fragments of sentences.

Perhaps the closest related work is that of Collins-Thomson and Callan [7]. They use a unigram language model to predict the reading difficulty level of a (long) text. Essentially their algorithm can predict the level (grade 1-12) that the text readabilty comes close to. However, we find that the correlation of their algorithm to the human judgments of the (short) web abstract is very weak.

In the web world, the summaries are very short, incom-

plete sentences that could have numerous non-alphabetic characters or exhibit strange capitalization. Therefore, the features that are important for the web summary are not likely to be included in FOG and other models. More recently a feature-based approach has been used to model answer quality [13, 2, 19]. Our work, however, focuses on the issue of modeling readability itself and using it for realtime monitoring of readability of web result summaries.

Researchers in psychophysics [24] have also studied various aspects of reading: impact of context on readability; eye movements during reading; and prediction based on context. Others [17] have looked at the impact of contrast and size on reading. These characterizations can help in creating new features and explaining why a feature is important.

Modeling of *response* variables (the human judgments) as a function of *explanatory* variables (e.g. measured readability features) has been studied under the names of statistical inference [31], pattern recognition [12] and more recently statistical machine learning [11]. Techniques such as logistic regression, support vector machines (SVMs), neural networks, and decision trees have been popular in the modeling community. SVMs in particular have been used recently to learn ranking functions [21, 14, 4] and have been shown to work well. Friedman's stochastic Gradient Boosting Decision Tree (GBDT) [8, 9] is a promising new approach that computes a function approximation by performing a numerical optimization in the function space instead of the parameter space. In fact, recent papers [18, 33] have shown that GBDT can outperform many competing machine learning techniques. We use GBDT as our modeling algorithm, which is discussed in the next section.

3. MODEL

In this section we describe our regression approach to modeling summary readability. In the next subsection we describe the readability features (explanatory variables) we computed for modeling the human responses. In the following subsection we provide an overview of the gradient boosting algorithm.

3.1 Features

Features can be critical in any modeling task. A good modeling algorithm can not help if the features lack information about the concept being modeled. The features that we used in our experiments are listed below.

1. FOG: This is a readability measure based on features such as average number of syllables per word. It is a fixed linear formula that is computed from features. The weights are fixed and were estimated for long essays [10, 20].

2. Flesch: This metric is similar to FOG [16].

3. Flesch-Kincaid: This is another long prose metric similar to FOG and Flesch [16].

4. Average Characters Per Word (CPWRD).

5. Average Syllables Per Word (SYLPWRD):

6. Percentage of Complex Words (PCMPLXWRDS): This is a feature used by Flesh.

7. Number of Snippets (NSNIP): If one tries to squeeze in too many fragments into an abstract, it looks choppy and unreadable.

8. Does the Abstract Begin with Ellipses (BELLIP): Do beginning ellipses impact readability?

9. Does the Abstract End with Ellipses (ELLIP): Do ending ellipses impact readability?

10. Capital Letters Fraction (CAPFRAC): If there is overcapitalization, it should hurt readability.

11. Punctuation Character Fraction (PUNCFRAC): If there are too many punctuation marks, most likely it is spam or some sort of non-text document.

12. Stop Word Fraction (STOPFRAC): We notice that spammers typically try to insert multiple occurrences of the keywords. Fraction of stop words is a surrogate for a real language model [30, 32].

13. Query Word Hit Fraction (HITFRAC): Readers are influenced by the presence or absence of terms from the query.

The above features are computed after removing any hypertext markup like bolding and then stripping all ellipses. The features were computed on the entire abstract and not on an individual snippet.

An interesting point to note is that it is very difficult to know which features are the important ones without manually observing examples. However, using the machine learning technique described in the next subsection we can compute the relative influence of the individual features. This will be discussed again in the experiments section.

3.2 Gradient Boosted Decision Trees

A basic regression tree $f(x)$, $x \in R^N$, partitions the space of explanatory variable values into disjoint regions R_j, $j = 1, 2, \ldots, J$ associated with the terminal nodes of the tree. Each region is assigned a value ϕ_j such that $f(x) = \phi_j$ if $x \in R_j$. Thus the complete tree is represented as:

$$T(x; \Theta) = \sum_{j=1}^{J} \phi_j I(x \in R_j), \tag{1}$$

where $\Theta = \{R_j, \phi_j\}_1^J$, and I is the indicator function. For a given loss function $\psi(y_i, \phi_j)$ the parameters are estimated by minimizing the the total loss:

$$\hat{\Theta} = \arg\min_{\Theta} \sum_{j=1}^{J} \sum_{x_i \in R_j} \psi(y_i, \phi_j). \tag{2}$$

Numerous heuristics are used to solve the above minimization problem.

A boosted tree is an aggregate of such trees, each of which is computed in a sequence of stages. That is,

$$f_M(x) = \sum_{m=1}^{M} T(x; \Theta_m), \tag{3}$$

where at each stage m, Θ_m is estimated to fit the *residuals* from the $m-1$th stage:

$$\hat{\Theta}_m = \arg\min_{\Theta_m} \sum_{i=1}^{N} \psi(y_i, f_{m-1}(x_i) + \phi_{j_m}). \tag{4}$$

In practice, instead of making adding $f_m(x)$ at the mth stage, one adds $\rho f_m(x)$ where ρ is a the *learning rate*. This is similar to a "line search" where one moves in the direction of the gradient, but the step size need not be equal to the gradient. In the *stochastic* version of GBDT, instead of using the entire dataset to compute to loss function, one sub-samples the data and then finds the function values ϕ_j such that the loss on the test set is minimized. The stochastic variant minimizes over fitting issues. The depth of the trees in each stage is another algorithm parameter of importance. Interestingly, making the trees in each stage very shallow while increasing the number of boosted trees makes good function approximation. In fact, even with depth 1 trees we get good results. Interaction amongst explanatory variables is modeled by trees of depth greater than 1.

Finally the GBDT algorithm also provides what is called *relative influence* of variables. This is computed by keeping track of the reduction in the loss function at each feature variable split and then computing the total reduction of loss function along each explanatory feature variable. The higher the relative influence value of a feature, the more important it is in the prediction process.

Ridgeway's article [26] provides a nice survey of boosting algorithms, and we use his GBM package [25] and R [1] for conducting our modeling experiments.

4. EXPERIMENTAL PROTOCOL

The training data for the model consisted of 5382 judgments of readability done by seven human editorial judges over about a year, on abstracts from Yahoo! and Google search results. Each result was rated on a scale of 1-5, where 1 was the least readable, and 5 was the most like written English. The editorial judgment tool is shown in Figure 2.

Editors were instructed to consider whether the text was written intended to be read as English, or whether it was from text on the page, such as menus, that was not part of the written content of the page. Elements of readability or unreadability are shown in Table 1.

This judgment task was part of a larger task in which editors were asked to also judge perceived relevance and the quality of certain kinds of semantic content in the abstract. One concern about judgments of this nature is that it is possible for different categories to be conflated, and for relevance to impact readability judgments. However, relevance judgments and readability judgments were only weakly correlated, with a correlation coefficient of 0.24. Because it is difficult to directly measure relevance in many real-world situations, we decided not to include relevance in our model, so as to produce a useful model in cases where we may not have relevance information.

The editors were also given examples of readable and unreadable abstracts, as shown in Table 2. The judgment training process included a number of prejudged cases, and editors were required to judge those correctly, as well as show good agreement with other editors on overlap judgments, where multiple editors judged the same abstract. Cases where the judgments disagreed by more than a single point were discussed and rejudged.

Inter-rater agreement was calculated for overlap judgments. Editors showed perfect agreement only 46.4% of the time, but showed near agreement, where near agreement is agreement to within one point on the 1-5 scale, 84.5% of the time. Editors showed perfect agreement with themselves on abstracts they rejudged 95.7% of the time, and near agreement 99.7% of the time.

In Table 3 we examples of causes that lead to low readability.

There are numerous parameters values that need to be set for the gradient boosting algorithm. These parameters were described in the section describing the model. In particular, the number of trees we used was set at 3000, the shrinkage factor was 0.05, the interaction depth was 2, and the bagging fraction was 0.5. Also, the minimum number of observation in any node was set at 10.

The training and test set were created by randomly splitting the 5382 observations into equal sized training and test sets. Finally, for model selection, we used 5 fold cross validation to estimate the model on the training set. The results reported are on the held-out test set.

The FOG, Kincaid and Flesch-Kincaid features were computed using the Fathom [28] package. We used R [1] and GBM [25] to do the statistical modeling.

5. RESULTS AND DISCUSSION

The purpose of the experiments was to

1. To see if there is any correlation between the predicted judgment values and the true human judgment.
2. Understand which features are relatively more important for predicting readability values
3. Investigate the nature of issues that lead to disagreements between the predicted and true judgment values.

The exact protocol used for the experiments are described in the previous section; here we discuss the results.

In Figure 3 we show a scatter plot of FOG, Kincaid and Flesch-Kincaid scores and human judgments. Interestingly there does not seem to be any correlation. This most likely is due to the fact that web abstracts i) have very little text (at most two lines); ii) are comprised of small fragments of sentences instead of complete sentences; and iii) the models are trained on a completely different corpus. The visual lack of correlation shown in Figure 3 is validated in Table 4 where we provide the Pearson's correlation coefficient for the three readability metrics. The Pearson's coefficient lies between 1 and -1 and the values corresponding to Fog, Kincaid, and Flesch-Kincaid are around 0, indicating the lack of correlation. We also computed the Pearson's correlation coefficient for the scores generated by the Collins-Thompson-Callan model and the value indicates negligible correlation. In fact, the Collins-Thompson-Callan model score mean was 7.76 with a standard deviation of 2.79, indicating that the on average the reading difficulty of web abstracts is at the level of a 8th grader. This in fact emphasizes the need for specialized models for short abstracts – the Collins-Thompson-Callan model expects larger texts whereas web abstracts are very short.

In Figure 4 we plot the predicted judgment score against the true human judgment score. The true human judgment scores are integral values from 1 to 5. The estimated score, however, is a continuous value, since the gradient boosting was used in the regression mode and not in the classification mode. Thus the fitted points are clustered vertically around the integers 1 through 5. The plot on the left is for

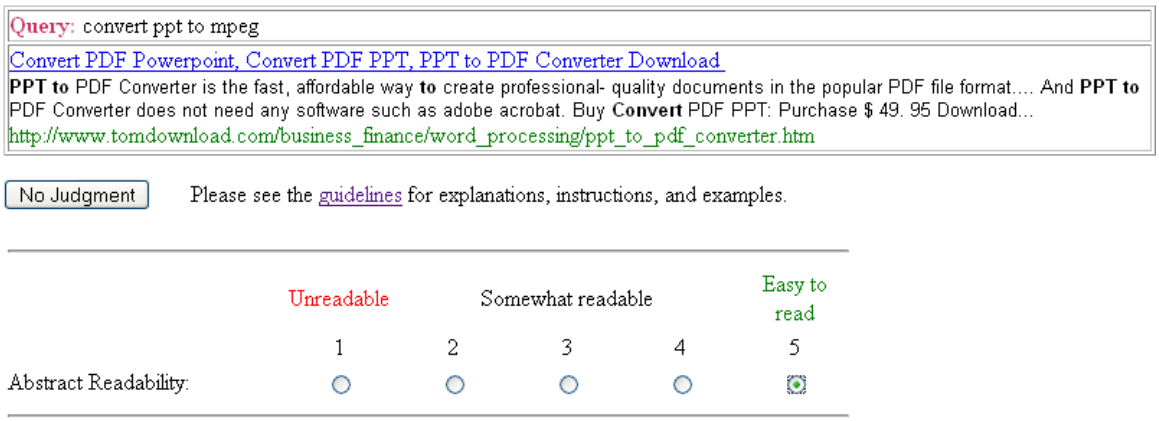

Figure 2: Tool provided to the editors to rate the abstracts.

Table 1: Elements of readability. Instructions provided to editors defining the notion of readability

Attributes of a readable abstract	Attributes of an unreadable abstract
it allows us to quickly scan and understand its gist	choppiness from keyword loading or content taken from navigational or menu lists.
snippets are portions of text clearly intended to be read by a human	snippets with poor truncation, e.g., broken at a prepositional phrase, conjunction, adjective/noun, etc.
snippets are generally complete sentences, coherent excerpts of sentences, or understandable titles	snippets that provide little information or query term context, because they are too short, the truncation is bad, etc.
it looks nice, without garbage characters, wingdings, all caps, etc.	snippets that contain many misspellings or are poorly written, perhaps written by a non-native English speaker

Table 4: Pearson correlation. We see that Web abstracts, which are comprised of short phrases, are not well correlated with the human readability judgment scores. In addition, the linear model can not predict the readability well. The scores predicted using gradient boosted decision trees, however, have a much better correlation with human readability judgments

Score Type	Pearson's coefficient
Fog	0.01572242
Kincaid	-0.02689905
Flesch-Kincaid	0.02323278
Linear	-0.001198311
Collins-Thompson-Callan	0.0597
Gradient Boosted Trees	0.6321157

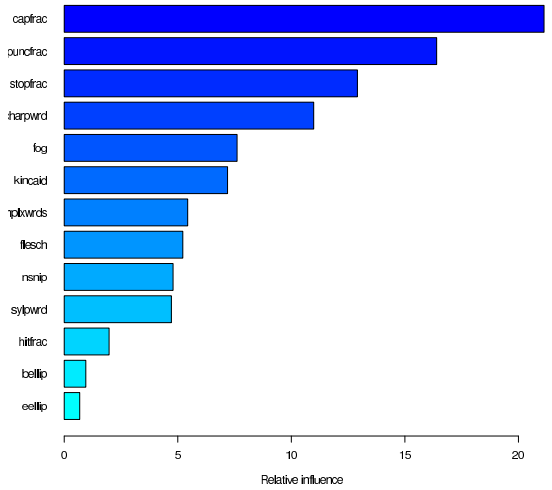

Figure 5: Relative influence of the features. It can be seen that CAPFRAC, the fraction of capital letters, is an important feature for predicting readability.

predictions using a linear model, and the plot on the right is for predictions using the gradient boosted decision trees. We see that GBDT has a much better correlation than the linear model. In Table 4 we give the Pearson's correlation coefficient for both the linear and GBDT models and it is clear that the GBDT model performs the best.

The next obvious question to ask is, where do the predictions and true values disagree the most. We will address this point shortly.

The second issue is that of relative influence (see modeling section for definition) of features in readability modeling. In Figure 5 we see that the top three most influential feature are CAPFRAC (fraction of capital letter), PUNCFRAC (fraction of punctuation characters), and STOPWRD (fraction of stop words). Notice that these features are not considered in the standard FOG and Flesch-Kincaid algorithms. CAPFRAC is important because when the abstract is made up of all capital letters, it makes the presentation very ugly. Too much punctuation behaves similarly. This can also be an indication of spam or junk text. Lack of stop words is a clear sign of spam, since spammers tend to insert multiple occurrences of key words in the text, often in a list.

Notice that CHARPWRD (average characters per word), which is a component of the standard FOG-type metric, is actually fourth in importance, higher than any of the FOG-type metrics. Finally, while the number of snippets

Table 2: Examples of readable and unreadable abstracts that were provided to the editors. We also specified the grades associated with the examples

Grade	Explanation	Example
1	"unreadable"	... the pulldown: Choose Video Here: Disclaimer Tread-Mill Got a Light? Quicktime Required. ...MMV New Line Productions, Inc. The Twilight Zone ...
3	"somewhat readable"	...Courtesy of Geffen Records. Favorites. Alerts. The new-school punk trio blink-182 was formed near San Diego, California around guitarist/vocalist Tom...
5	"easy to read"	... a better way to buy diamonds and engagement rings at the best prices on the web ... Email us the carat weight, clarity, color, shape and price range of what diamonds you want, and we will search the main wholesale market for you

Table 3: Common readability errors

Lack of context for bolded query term	Query term occurs without sufficient context, for example, as the last word of a snippet.
Broken prepositional phrase	Snippet starts/ends in the middle of a prepositional phrase, e.g. "do not use this product with..."
Broken at conjunction	Snippets starts/ends with a conjunction, e.g. "Tonight I'm going to the store and..."
Broken adjective/noun	Snippet starts/ends between adjective and noun or noun and determiner, e.g. "make sure to use new..."
Choppy or unreadable snippet	Snippet does not form a connected, readable block of text, e.g. ".. the pulldown: Choose Video Here: Disclaimer Tread-Mill Got a Light? Quicktime Required. _ & MMV New Line Productions, Inc. The Twilight Zone ..."
Split married word	Snippet starts/ends between parts of a phrase that has a coherent meaning together ("ice cream," "minimum wage," "New York," etc.) e.g. "My favorite dessert is ice..." or (on a page about the band Cheap Trick) "One of the most popular bands of the late 1970s was Cheap..."
Broken subject/verb or verb/object	Snippet starts/ends between subject and verb or between verb and object, e.g. "...broke the record for 100 meters."
Broken adverb/verb	Snippet starts/ends between adverb and verb, e.g. "After careful planning, the thieves brazenly..."
Broken address	Snippet starts/ends in the middle of an address, e.g. "701 First Avenue..." (no city)
Broken phone number	Snippet starts/ends in the middle of a phone number, e.g. "408 349..."
Broken date/time	Snippet starts/ends in the middle of a date and/or time, e.g. "August 29,..." (no year)
Garbage characters, dingbats, etc.	
Navigational, menu, lists in snippets	e.g. "My eBay. My eBay Views. My Summary. All Buying. Bidding. Won. Didn't Win. My Messages. All Favorites. My Account. Related Links."
JavaScript code in snippets	This category applies to any text intended to be read by the computer, not by a human – CSS, raw HTML, etc. e.g. "var PUpage="76001078"; var PUprop="geocities";yfiEA(0);geovisit();

(NSNIP) is not the lowest in the list, it is less important than CAPFRAC.

The machine learning methodology thus provides an easy way to not only predict the judgments but also provide an explanation of which features are important in decision making.

In Figure 6 we provide a number of cases where the model agreed with human judges that the readability of the abstracts is of low quality. In Figure 7 we provide a number of cases where the model disagreed with the human judges – the judges rated the abstract as of low quality whereas the algorithm rated them as of high quality. In some cases we see that, although the abstract is quite readable, the judge gave it a low score, perhaps due to objectionable content.

In Figure 8 we provide a number of cases where the model agreed with human judges that the readability of the abstracts is of high quality. In Figure 9 we provide a number of cases where the model disagreed with the human judges – the judges rated the abstract as of high quality whereas the algorithm rated them as of low quality. There are cases where we see that that while one of the snippets was very unreadable, the human judge still gave the abstract a high score, focusing on the readable snippet only.

This analysis leads us to believe that one can improve the guidelines to make the features and models better reflect how human judges score abstracts. It also suggests that instead of making features aggregate values over the entire abstract, perhaps the features should be on a per snippet basis.

Once a readability model is estimated as described in this article, it can be used for monitoring the readability of abstracts viewed by users on a daily basis instead of the usual quarterly or yearly evaluation using human judgments. At the same time it is important to emphasize that the predictions are only surrogates for the real human judgements. Finally, while the proposal is for computing readability of abstracts in real time and at a very large scale, one can use the same features to design summarizers that produce more readable abstracts.

6. CONCLUSIONS

While TREC-style evaluations are very valuable for the research community, the process itself does not lend itself for conducting real time quality evaluation of search engine summary results. In this paper we present a machine-learning methodology that first models the readability of abstracts using training data with human judgments, and then predicts the readability scores for previously unseen documents using gradient boosted decision trees.

The performance of our model exceeds that of other kinds of readability metrics such as Collins-Thompson-Callan, Fog or Flesch-Kincaid. This is not surprising since these models

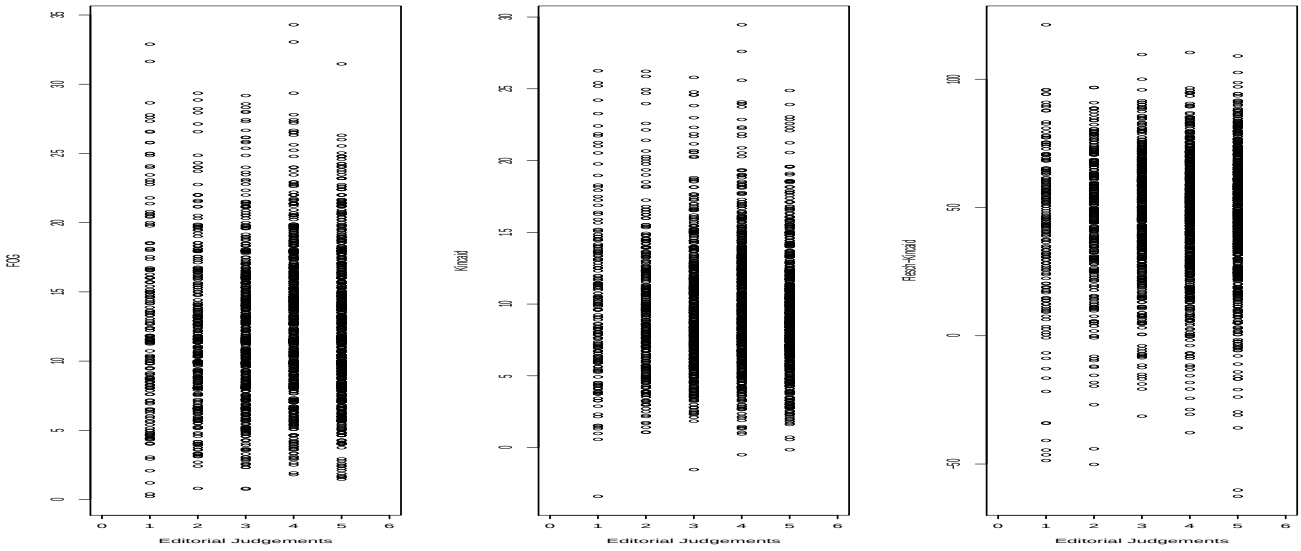

Figure 3: Correlation between FOG, Kincaid and Flesch-Kincaid readability metrics. On the x-axis we have human readability judgments from 1-5 and on the y-axis we have the estimated readability metric scores. We can see that there is hardly any correlation, which is quantified using the Pearson correlation coefficient in Table 4.

were trained on clean text samples, with complete sentences, and larger text sizes. In contrast, web abstracts can contain many non-standard characters and they typically have fragments of sentences.

Our methodology can be used to estimate human quality judgments in real time and at a large scale, and is being used for that purpose now. Furthermore, the model can also be used in the automatic summarization algorithm to generate more readable summaries.

7. ACKNOWLEDGMENTS

Kanungo would like to thank Jamie Callan for making a web-based version of his readability system available to us.

8. REFERENCES

[1] The R project for statistical computing. `http://r-project.org`.

[2] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *Proc. of WSDM*, 2008.

[3] A. Aula. Enhancing the readability of search result summaries. In *Proc. of HCI*, 2004.

[4] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *Proc. 22nd Proc. Intl. Conference on Machine Learning*, pages 89–96, 2005.

[5] J. Burstein, K. Kukich, S. Wolff, C. Lu, M. Chodorow, L. Braden-Harder, and M. D. Harris. Automated scoring using a hybrid feature identification technique. In *Proc. of the 17th Intl. Conference on Computational Linguistics*, 1998.

[6] C. L. A. Clarke, E. Agichtein, S. Dumais, and R. W. White. The influence of caption features on clickthrough patterns in web search. In *Proc. 30th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 135–142, 2007.

[7] K. Collins-Thompson and J. Callan. A language modeling approach to predicting reading difficulty. In *Proceedings of HLT/NAACL*, 2004.

[8] J. H. Friedman. Greedy function approximation: A graidient boosting machine. *Annals of Statistics*, 29:1189–1232, 2001. `http://www-stat.stanford.edu/~jhf/ftp/trebst.pdf`.

[9] J. H. Friedman. Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38:367–378, 2001. `http://www-stat.stanford.edu/~jhf/ftp/stobst.pdf`.

[10] R. Gunning. *The technique of clear writing.* McGraw-Hill, 1952.

[11] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning.* Sringer-Verlag, New York, NY, 2001.

[12] A. K. Jain, R. P. W. Duin, and J. Mao. Statistical pattern recogntion: A review. *IEEE Transactions on Pattern Analysis and Machine Learning*, 22:4–37, 2000.

[13] J. Jeon, B. W. Croft, J. H. Lee, and S. Park. A framework to predict the quality of anwers with non-textual features. In *Proc. of 29thAnn. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 228–235, 2006.

[14] T. Joachims. Optimizing search engines using clickthrough data. In *Proc. 8th Ann. Intl. ACM SIGKDD Conf. on Knowledge Discovery and Data Mining*, pages 133–142, 2002.

[15] M. D. Kickmeier and D. Albert. The effects of scanability on information search: An online experiment. In *Proc. of HCI*, 2003.

[16] J. P. Kincaid, R. P. Fishburn, R. L. Rogers, and B. S.

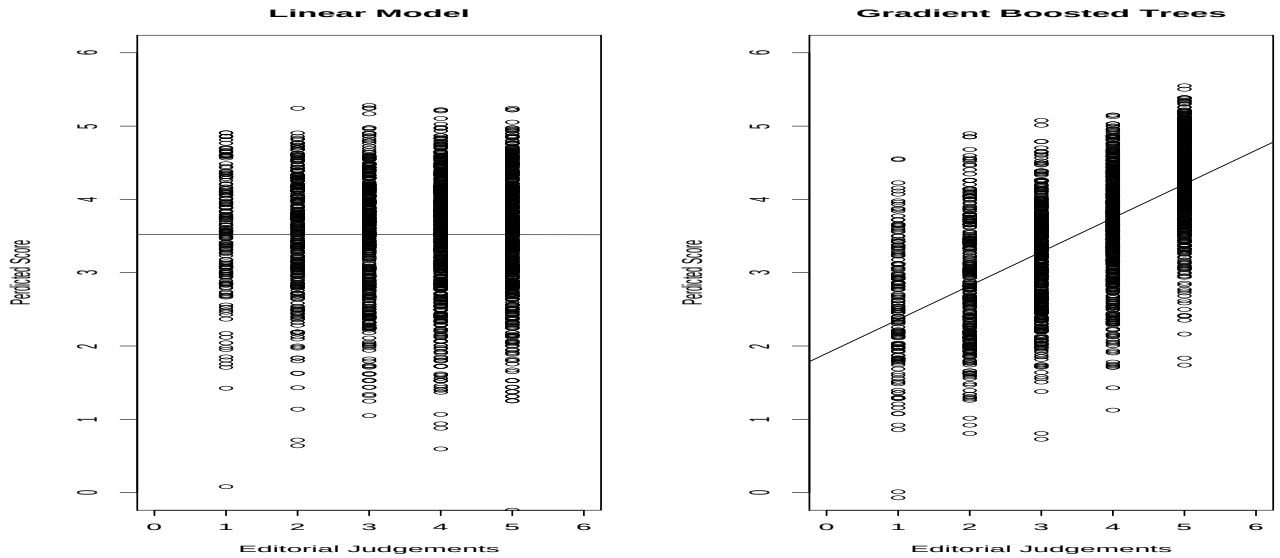

Figure 4: True versus predicted judgments, using i) linear and ii) gradient boosted decision trees. On the x-axis we have human judgments from 1-5 and on the y-axis we have the predicted judgments. Since the human judgments are integer values from 1-5, the points are clustered vertically around integers. Regression using linear and gradient boosted decision trees produce non-integral scores and hence the y-values are not integral.

Chissom. Derivation of new redability formulas for navy enlisted personnel. Technical report, Milington, Tenn, Naval Air Station, 1975. Tech Report Research Branch Report 8-75.

[17] G. Legge. *Psychophysics of Reading in Normal and Low Vision*. Lawrence Erlbaum Associates, 2006.

[18] P. Li, C. J. Burges, and Q. Wu. Mcrank: Learning to rank using multiple classification and gradient boosting. In *Proc. 21st Proc. of Advances in Neural Information Processing Systems*, 2007.

[19] S. F. Liang, S. Delvin, and J. Tait. Evaluating web search result summaries. In *European Conference in IR Research*, pages 96–106, 2006.

[20] G. H. McLaughlin. SMOG grading: A new readability formula. *Journal of Reading*, 12:639–646, 1969.

[21] R. Nallapati. Discriminative models for information retrieval. In *Proc. 27th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 64–71, 2004.

[22] H. Obendorf and H. Weinreich. Comparing link marker visualization techniques: Changes in reading behavior. In *Proc. of 12thIntl. Conference on the World Wide Web*, pages 736–745, 2003.

[23] D. R. Radev and W. Fan. Automatic summarization of search engine hit lists. In *Proc. of ACL*, 2000.

[24] K. Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124:372–422, 1998.

[25] G. Ridgeway. Generalized boosted models: A guide to the gbm package. `http://i-pensieri.com/gregr/papers/gbm-vignette.pdf`.

[26] G. Ridgeway. The state of boosting. *Computing Science and Statistics*, 31:172–181, 1999. `http://www.i-pensieri.com/gregr/papers/interface99.pdf`.

[27] D. E. Rose, D. M. Orr, and R. G. P. Kantamneni. Summary atributes and perceived search quality. In *Proc. of Intl. Conference on the World Wide Web*, 2007.

[28] K. Ryan. Fathom. `http://search.cpan.org/dist/Lingua-EN-Fathom`.

[29] L. Si and J. Callan. A statistical model for scientific readability. In *Proc. of the 10th Intl. Conference on Information and Knowledge Management*, 2001.

[30] F. Song and W. B. Croft. A general language model for information retrieval. In *Proc. 8th Intl. Conf. on Information and Knowledge Management*, pages 316–321, 1999.

[31] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Sringer-Verlag, New York, NY, 2002.

[32] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proc. 24th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 334–342, 2001.

[33] Z. Zheng, H. Zha, T. Zhang, O. Chapelle, K. Chen, and G. Sun. A general boosting method and its application to learning ranking functions for web search. In *Proc. 21st Proc. of Advances in Neural Information Processing Systems*, 2007.

Human Score	Predicted Score	Query, Title, Abstract, URL, Comment
1	1.1	Query: bad day lyricks Title: R.E.M. - Bad Day Lyrics / Songteksten / Songtexten / Album Covers Abstract: ... bad day r.e.m. songtext bad day lyric bad day lyrics bad day ... nummer bad day songtekst bad day songteksten song bad day r.e.m. tekst bad day lyricks bad day text ... URL: http://www.sleurink.nl/song/7916/lyrics/rem/bad-day-songtekst.html Comment: Good prediction
1	1.11	Query: +laptop +repair +ompaq Title: new year price reductions laptop price reductions new laptop ... Abstract: ... Low-cost data recovery service.....7 Dell laptop repair .....7 ... A 5.5 / 2.5 H ol ow Ã‚Â£ 49 AC-19V-120W-C OMPAQ 18-20v ... URL http://www.portables.co.uk/PriceList.pdf Comment: Good prediction
1	1.22	Query: asicscheerleading shoes Title: Cheerleader Shoe Charm Abstract: ... cherleader cheelreading sohe uresics cheersalader asicscheerleading cheerleoding sharm cheerleyading cheerleiding cheerl cheerleauder ... URL: http://www.totalfitnessfl.com/fit/Cheerleader_Shoe_Charm_z42000_66832.php Comment: Good prediction

Figure 6: Examples of summaries that were assigned low readability scores by human judges, and low readability scores by the algorithm.

Human Score	Predicted Score	Query, Title, Abstract, URL, Comment
1	4.5	Query: cool website Title: StatCounter Free invisible Web tracker, Hit counter and Web stats Abstract: StatCounter Free invisible Web tracker, Hit counter and Web stats StatCounter Free invisible Web tracker, Hit counter and Web stats A free yet reliable invisible web tracker, highly configurable hit counter and real-time detailed web stats... URL: http://www.statcounter.com/ Comment: Bad prediction. Notice how part of the abstract is listy and the rest is good English.
1	4.4	Query: strip club las vegas Title: The Ultimate Guide to Las Vegas Strip Clubs Abstract: ... the best possible product to those looking for a good strip club in Las Vegas. I decided to enlist a very good friend of ... URL: http://govegas.about.com/cs/adultinterests/a/Stripclubguide.htm Comment: In this case perhaps the human judgment is incorrect; maybe influenced by the topic.
1	3.97	Query: seamlessweb Title: SeamlessWeb: Login (Order Food Online) Abstract: to the office, then SeamlessWeb Business Solutions is right for you. Username. Password ... best restaurants in the neighborhood, let SeamlessWeb take your order. Sign Up Here. Email Address. Select Your City ... URL: http://www.seamlessweb.com/ Comment: Bad Prediction -- listy and choppy.

Figure 7: Examples of summaries that were assigned low readability scores by human judges, and high readability scores by the algorithm.

Human Score	Predicted Score	Query, Title, Abstract, URL, Comment
5	5.3	Query: www.ksbe.edu Title: Kamehameha Schools - Contact Information Abstract: Founded in 1887, Kamehameha Schools is a statewide educational system supported by a trust endowed by Princess Bernice Pauahi Bishop URL: http://www.ksbe.edu/contact.php Comment: Good Prediction; regression is not constrained to be integers.
5	5.26	Query: firehouse Title: Firehouse.com Abstract: News and information for the emergency services industry. URL: http://www.firehouse.com/ Comment: Good Prediction
5	5.18	Query: fantasy baseball Title: Yahoo! Sports Fantasy Baseball Abstract: Compete with others playing fantasy baseball online. Free and pay leagues available. URL: http://baseball.fantasysports.yahoo.com/b1 Comment: Good Prediction

Figure 8: Examples of summaries that were assigned high readability scores by human judges, but low readability scores by the algorithm.

Human Score	Predicted Score	Query, Title, Abstract, URL, Comment
5	1.8	Query: nascar.com Title: NASCAR.com : races : tracks : Atlanta Motor Speedway Abstract: Track Facts. Banking/Turns: 24. Distance: 1.54 miles. Shape: Oval. Nextel Cup Race Record. Bobby Labonte 159.904 11116l97. Busch Race Record. Mark Martin 151.751 03108l97. Nextel Cup Qualifying Record URL: http://www.nascar.com/races/tracks/ams Comment: Perhaps human judgment not correct -- too listy.
5	2.01	Query: eva longoria Title: Eva Longoria Abstract: Eva Longoria - Filmography, Awards, Biography, Agent, Discussions, Photos, News Articles, Fan Sites. URL: http://www.imdb.com/name/nm0519456/ Comment: Perhaps human judgment not correct, or guidelines need to be modified. Listy but informative.
5	2.23	Query: oprah winfrey Title: Oprah's Angel Network Abstract: Founded by Oprah Winfrey. Organization supporting programs for women, children, and families, educational programs, and health and human services. ... Oprah. SEE WHAT'S NEW IN THE ANGEL NETWORK. Africa Grants ... URL: http://www.oprah.com/uyl/uyl_landing.jhtml Comment: An all captital snippet contributed to the the abstract being ranked low; but human judge ignored it. Guidelines issue?

Figure 9: Examples of summaries that were assigned high readability scores by human judges, and high readability scores by the algorithm.