Exploring the Feasibility of Automatically Rating Online Article Quality

Laura Rassbach
University of Colorado
Department of Computer Science
Boulder, CO 80309-0430

Trevor Pincock
University of Colorado
Department of Linguistics
Boulder, CO 80309-0430

Brian Mingus
University of Colorado
Department of Psychology
Boulder, CO 80309-0430

ABSTRACT

We demonstrate the feasibility of building an automatic system to assign quality ratings to articles in Wikipedia, the online encyclopedia. Our preliminary system uses a Maximum Entropy classification model trained on articles hand-tagged for quality by humans. This simple system demonstrates extremely good results, with significant avenues of improvement still to explore.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

Quality, Wikipedia, Maximum Entropy

1. INTRODUCTION

The quality of a literary endeavor is often the first and last thing anybody needs to know about it. People are generally very good at identifying high-quality writing, though some are better than others. The quality of writing is related not only to a piece's readability, but also to its accuracy and informativeness. The subjective nature of quality makes computers very bad at deciding quality, to the delight of the mediocre blogger and the text-generating spam bot. As the amount of electronic data available on the Internet increases, and high-quality writing continues to be overwhelmed by low-quality work, automatic classification of quality becomes more important.

Wikipedia is an Internet-based, free-content encyclopedia. With 1.7 million English articles, Wikipedia dwarfs its next largest competitor, the venerable Encyclopedia Britannica with 120 thousand articles. Officially launched in 2001, Wikipedia's rapid exapansion to become the world's largest English enclyclopedia is remarkable. The size and growth of Wikipedia is due to the efforts of a volunteer army of contributors and editors. Wikipedia's free content policy allows anyone to edit or create articles. This laissez-faire attitude encourages contribution, but also allows poor quality articles, heavily biased and containing misinformation to enter the encyclopedia. The editorial team at Wikipedia monitors the content, but the enormity of the task makes errors inevitable. Wikipedia's founder, Jimmy Wales, has admitted that quality control is an important problem for Wikipedia to address [13].

Despite this weakness, Wikipedia's accuracy has stood up to review. A peer review of Wikipedia and the Encyclopedia Britannica's scientific articles yielded no significant difference in error averages per article between the volumes [6]. Usage of Wikipedia has reached astonishing levels as well. According to a poll by the Pew Research council over 30% of Internet users utilize Wikipedia as a educational resource. The usage trend increases with educational level, as 50% of Internet users with a college degree use the online encyclopedia. Because of its enormous size, Wikipedia is also increasingly becoming a valuable resource in Natural Language Processing. It has been used in tasks such as word sense disambiguation, co-reference resolution, and information extraction [20, 14, 21]. The availability of quality ratings for Wikipedia articles would assist both human users and automatic applications in selecting the best articles for their purposes.

Quality is notoriously hard to quantify [19]. For a multimedia entity, such as Wikipedia, overall quality is the composition of the various parts. Currently there are computational applications which are capable of assessing the quality of some of the components. The GRE utilizes a program to evaluate essays in conjunction with human evaluations [15]. Text in encyclopedic articles does not exactly match that of an essay, but many of the principles are the same (for example, clear and well-written prose is important in both cases). There exist no established criteria for evaluating the quality of other elements within an article, such as the images, time lines, topical hierarchy, citations, and others. Although it is difficult to determine quantitative measures of quality, it is easy for people to determine the relative quality of something. A subset of Wikipedia articles have been assessed and annotated by the community of users. We used the Maximum Entropy machine learning technique to train a classifier on this dataset to automatically evaluate the quality of articles. Despite a limited number of features, we have obtained significant results. The machine learning approach

can reverse-engineer the quality assessments of human annotators. We suggest features to extend the classifier and applications of the research.

2. RELATED WORK

Content creators will necessarily be concerned with quality. Simple extensions of applications that help creators ensure quality become popular immediately. The spell check feature of word processors has saved many papers from embarrassing errors and sounded the death knell for typewriters that didn't perform this basic quality assurance operation. The grammar check feature of some word processors attempts to expand on this very useful tool by incorporating syntactic rules to aid composition. However, the conventions of grammar are more mutable than those of spelling, making these systems error prone. Although these systems are not perfect research indicates that students using word processors produce higher quality work than those who do not [1]. The Writer's Workbench was a program developed in the 1980s that detected things such as split infinitives ("to boldly go"), overly long sentences, wordy phrases and passive sentences [12]. These metrics of quality are relatively objective. The features that make up great writing extend beyond sentence boundaries, are inherently subjective, and are much harder to evaluate. Grammar models that rely on rules or statistical regularity would balk at Shakespeare's work.

Discourse analysis is typically concerned with measurements of cohesion and coherence. Cohesion refers to the relationship between lexical units; coherence is the relationship between the meaning of units of text. Simple measures of cohesion would capture some of the nuances of discourse quality. The TextTiling algorithm proposed by Hearst measures cohesion by segmenting the discourse and measuring the lexical similarity between segments [8]. Latent Semantic Analysis, an algorithm which primarily measures the similarity of documents by word frequency, is another way to measure cohesion. Foltz et al describe using LSA to measure the quality of student essays [4]. Their results indicated that LSA could be used to achieve human accuracy in holistic judgments of quality. The limitation of LSA is that the domain must be well defined and a representative corpus of the target domain must be available.

Witte and Faigley offer a critique of using cohesion as a measurement of writing quality [26]. They argue that writing quality cannot be divorced from context, and factors such as the writer's purpose, the discourse medium and the characteristics of the audience are essential to qualitative analysis. While the presence or the absence of cohesion doesn't confirm or disconfirm quality, it is a useful indicator.

Coherence is a more reliable indicator of quality, but coherence is more difficult to quantify. Centering Theory presents a model of how coherent discourse should be structured. The theory posits that a discourse focuses on a single entity and that all utterances are centered on the entity and the introduction of new objects of focus must be done in relation to the centered objects, and it defines criteria for these transitions and ranks them in preferential order [7]. Miltsakaki and Kukich applied Centering Theory's hypothesis of attentional shifting to essays evaluated by Educational Testing Services' e-rater essay scoring system [15]. They found that the number of Rough-Shifts correlated with a lower score from e-rater. Their dataset had to be hand annotated to represent the roles of constituents in Centering Theory, making such analysis time consuming.

3. METHODOLOGY

3.1 The Dataset

The Wikipedia Editorial Team has begun tagging articles according to their quality. Articles are assigned to one of six quality classes: Featured, A, Good, B, Start, and Stub. Hand annotation is slow, laborious work. Assessments are based on the judgments of the annotators and not a quantitative analysis. There are established criteria for article classification, defining what articles of particular quality should be like. Featured Articles should be "well written, comprehensive, factually accurate, neutral and stable." [9]. Definitions are given for each of these properties, but the inherent subjectivity of the evaluations is obvious. Nearly 600,000 articles have been tagged [10]. The vast majority of articles, 71%, have been classified as Stubs, or articles of very short length containing incomplete material. Most articles begin as stubs and await further content contributions. Stub articles are relatively easy to recognize because of their brevity. However, they aren't very useful for a categorization task of quality, because they lack many of the elements of the more complete articles which are suitable as an educational resource.

The remaining data set of rated articles, with Stub articles removed, contains 168,183 total articles. The ratings for the articles that make up this set are as follows: 132,146 Start (78.6%), 31,600 B (18.8%), 2132 GA (1.3%), 873 A (0.5%), and 1432 FA (0.9%). The distribution is clearly skewed to the lower quality articles, with very few examples of articles the Wikipedia Editorial Team considers to be of "publishable quality."

We processed the rated articles to use in the training and testing of our classifier. The articles required quite a bit of preprocessing before they could be analyzed by our algorithms. We created separate entries in the database for HTML and plain text versions of the articles. The plain text of the articles was acquired using a Python module called BeautifulSoup [23]. The text was segmented into sentences by using MxTerminator, a Java implementation of a maximum entropy model specifically trained for sentence boundary detection [22].

3.2 Maximum Entropy Model

A Maximum Entropy (MaxEnt) model is a supervised machine learning algorithm used for classification, equivalent to a statistical regression algorithm. The algorithm uses a set of manually defined features to attempt to determine the probability of each example being in each class. The term "Maximum Entropy" refers to the fact that this classification is done making a minimum number of assumptions. For example, if the classifier has seen that 50% of training examples with feature 'x' are in class 'A', and has no other information, it will guess that the probability of a new example with feature 'x' being in class 'A' is 50%. Since it

has no other information, the rest of the probability mass will be distributed evenly among the other classes, since to do otherwise would make an assumption about the remaining classes that is not justified from the training data. A MaxEnt classifier works by assigning weights to each feature for each class, the list of features and classes are taken from the training data. For each class, each feature is multiplied by the associated weight and summed with the other features, then normalized to obtain the probability of the example being in the class. Features that interact, for example, 'number of words per paragraph' is an interaction between 'number of words' and 'number of paragraphs', must be manually combined into a single feature; the MaxEnt algorithm does not automatically analyze any interactions between features. The model learns by adjusting these weights iteratively based on the examples in the training set. [11, 3] There are several good reasons to use a MaxEnt classifier for this problem. MaxEnt classifiers are relatively simple and converge quickly, allowing us to experiment with adding new features to see their effect on the classification accuracy. At the same time, they are powerful systems that can succeed at quite difficult problems [16]. Finally, a MaxEnt model is built around the assumption that a human expert can easily identify all of the features likely to give clues about the correct classification of each example. This is an elegant approach to the quality classification problem because the original quality rankings are based on features observed by human experts – in other words, we believe that a set of features is already in use for hand classification, so it makes sense to attempt to use those features for automatic classification [5].

For our system, we used a Maximum Entropy classifier written in C++ with a Python wrapper by Zhang Le [27]. This classifier has a number of features useful for our problem. Unlike many implementations of MaxEnt, it allows the definition of non-binary features. This is convenient because it allows us to enter features such as length as the actual values rather than having to artificially decompose the values into a set of binary features. Our classifier uses the Limited-memory Broyden-Fletcher-Goldfarb-Shanno quasi-Newton algorithm to estimate the model weights [17]. We also use Gaussian Prior Smoothing with a variance of 2 to avoid overfitting the training set.

To train the MaxEnt model, we randomly selected an equal number of pages from each Wikipedia quality category (the actual number of selected pages is 650, 80% of the 'A' class articles). It is important to select an equal number of pages from each category because the model is powerful enough to use the distribution of the training data to assist in classification. Since the 'Start' category is overwhelmingly common, a model trained on a set with the same distribution as the actual page set will simply assign 'Start' to every page, and, having found a local maximum for classification, will cease to examine other features to improve the classification accuracy. Research has shown that altering the distribution of a training set for a machine learning algorithm often improves the performance of the algorithm on the test set [25]. We have artificially created a training set with equal numbers of examples in each category to force the classifier to learn the correct weights for the features we have defined. The weights for our model converge in less than 1000 iterations on this training set.

3.3 Feature Set

MaxEnt classifiers are typically built with extremely large feature sets, often as many as 10,000 features [24, 2]. Due to time constraints and equipment difficulties, we have an extremely limited set of only about 50 features. Our features fall into four general categories: length measures, part-of-speech usage, web-specific features, and readability metrics. Length measures include counts of the number of paragraphs and number of words, and give some hint as to whether the article is complete and comprehensive. Part-of-speech usage measures are counts of particular parts of speech in a syntactic parse of the article. These metrics allow us to begin to analyze the complexity of sentences and the quality of the article's prose. Web-specific features such as number of images and internal links reflect the authors' use of all the resources available for an article, as well as improving ease of understanding for readers. Finally, we used a number of standard readability metrics, including Kincaid, Coleman-Liau, Flesch, Fog, Lix, SMOG, and Wiener Sachtextformel, as another method of measuring the comprehensibility and complexity of an article's prose. These simple features begin to capture many of the qualitative assessments of the Wikipedia Editorial Team. For efficiency reasons, each feature was pre-computed and entered in the database, allowing us to add features and retrain the classifier relatively quickly.

4. RESULTS

A reasonable baseline classification system is one that classifies every article as a 'Start' article, since the overwhelming majority of articles in our dataset are 'Start' articles. This gives a classifier with a 78.6% accuracy. Despite our limited number of features and constraint of the training set to an equal distribution, our current classifier is nearly as accurate as the baseline, at 74.6% accuracy by article (that is, 74.6% of articles in the test set are classified correctly). Interestingly, Start-class articles are by far the least commonly mis-classified, probably because our feature set is most applicable to distinguishing Start articles from other classes. Our accuracy for the other classes is much lower, so that the normalized accuracy of our model is much lower, just under 50% (the normalized accuracy of the baseline model is 20%). However, this problem is at least in part because the other categories are much less distinct than the 'Start' category. If we collapse the category set into three categories instead of five ('Great', containing ratings F and A; 'Good', for rating G; and 'Poor' containing B and G) we see a normalized accuracy of 69% and a non-normalized accuracy of 91%. From the collapsed matrix we can see that 'Good' articles are the hardest to classify, probably because they have many of the characteristics of both 'Great' and 'Poor' articles. We expect to significantly improve the model's accuracy as we add more features, especially for this category.

5. FUTURE WORK

We've demonstrated the feasibility of classifying Wikipedia articles, and with modest improvements we could increase the accuracy of our system. We plan on expanding our classification system to include more features. More Wikipedia-specific analyses should improve performance. Additions to

Table 1: Confusion matrix for all Wikipedia quality categories. Categories along the top are the human-assigned rating; along the side are the ratings assigned by our classifier

		Correct Class				
		F	*A*	*G*	*B*	*S*
Classified As	*F*	0.44770	0.17391	0.20573	0.04835	0.00457
	A	0.16109	0.37267	0.14952	0.09526	0.01860
	G	0.20084	0.14286	0.41251	0.10088	0.02205
	B	0.11297	0.18012	0.14316	0.42347	0.12483
	S	0.07741	0.13043	0.08908	0.33204	0.82995

Table 2: Confusion matrix for collapsed categories. Categories along the top are the (compressed) human-assigned ratings; along the side are the (compressed) ratings assigned by our classifier

		Correct Class		
		Great	*Good*	*Poor*
Classified As	*Great*	0.593114	0.205726	0.046843
	Good	0.186228	0.562036	0.037549
	Poor	0.220657	0.232238	0.915608

the text analysis, such as more detailed analysis of syntactic structure, cohesion and coherence would help our system distinguish between low and high quality article better. Other applications, such as word processing, article retrieval, text summarization and spam detection, would benefit from automatic classification of quality.

Wikipedia articles contain a great deal of idiosyncratic formatting and information. A more thorough analysis of the features of Wikipedia's layout and wiki syntax would help to correctly classify articles. There are thousands of templates available for usage in a given article, and the number of templates used is strongly correlated with article quality. Features that assess the usage of a template would ensure that templates are used appropriately. The categorical organization of Wikipedia also allows for domain-specific analysis, allowing for a more disciplined analysis of word choice, style, and coverage. Every Wikipedia article also contains a history which stores all the edits made to it. The size of the Wikipedia history is roughly 30 times the size of the articles alone. The history would provide information about the creative process behind an article, and articles with more comprehensive histories would be assumed to be of higher quality.

Images are a strong indicator of quality. A colorful, informative illustration can elucidate a difficult concept, but a poorly chosen picture can actually detract from clever prose. Assessing the quality of an image computationally is a difficult endeavor. Judgments of picture quality are based on optical information and context, which is not readily available to a computer. The resolution of an image and the size can give a coarse idea of quality, as can a simple count of the number of images in an article. Digital photographs come with EXIF data that contains information about about camera settings and scene information that could also be relevant. Diagrams and explanatory figures would be more difficult to evaluate and a feature that merely detects their presence might be the most useful.

The PageRank Algorithm, formulated by Brin and Page [18], would be an excellent indicator of quality. The algorithm could be implemented using Wikipedia's internal link data. The more pages which link to a page would be indicative of its importance and indirectly of its quality. A more comprehensive implementation would incorporate pagerank information from the entire web; sites external to Wikipedia linking to a Wikipedia page would certainly suggest the article was of high quality. We are currently in the process of implementing this algorithm.

More sophisticated measures of the text will require additional parses including part of speech tagging, syntactic parsing, and dependency parsing. Such operations are computationally expensive, which would limit the applications of the system. Nonetheless, we plan to implement these features to assess their relevance to the quality of an article.

Our system is currently very good at distinguishing Start articles from all others. This isn't surprising considering the distribution of our dataset. Improving performance on the classification of the higher quality articles will entail differentiation between prose that is clear and grammatically correct and prose that is brilliant. This will require substantial discourse analysis.

We would also like to experiment with using a Support Vector Machine (SVM) for the classification task instead of the Maximum Entropy model. SVMs are similar to MaxEnt models in that they require the explicit definitions of features believed to give clues about the correct classification of examples. However, in contrast to MaxEnt, an SVM does not need interacting features to be explicitly defined. Rather, an SVM experiments with all possible feature combinations during training in order to discover combinations of features that allow an improvement in accuracy [11]. SVMs are often more accurate than MaxEnt models, but take significantly longer to train. In addition, they are often considered less elegant because the combinations they discover could have easily been added by hand, and it seems unreasonable and inefficient to attempt to automatically discover information we already know [5].

Many domains would benefit from automatic quality classification, particularly of text. Within Wikipedia, the automatic system could provide input to users as they are editing an article, suggesting areas of improvement. If the classifier was used on all articles, quality analysis by category would be more complete. Quality assessments of text could also be used for pedagogical purposes, to assist student's writing and provide instantaneous feedback and suggestions in an objective manner. There is the possibility of gaming such a system, but likely the things that would improve a quality rating would also improve the quality of a piece of text. Quality analysis would help in spam detection: most spam bots use automatic text generation, creating poor-quality, incoherent messages. Of course, this could just lead to more eloquent spam messages.

6. CONCLUSION

Classifying the quality of Wikipedia articles is an important task, since it can focus community attention on articles that need the most improvement and direct users to the articles most likely to be correct and informative. We have demonstrated that with minimal features a Maximum Entropy model can do a surprisingly good job of automatically classifying Wikipedia articles by quality. Our current model has an accuracy of 74.6%, which leaves room for improvement, but also shows the problem to be tractable. We enumerated a number of features that would enhance the model's performance.

7. ACKNOWLEDGEMENTS

Wikipedia runs on the custom-built MediaWiki platform, written in PHP and running on top of the MySQL database engine. All of Wikipedia is available for download in various formats, including XML, SQL, and HTML. The XML dump includes both embedded wikitext and metadata. The compressed archive containing all current versions of articles and content is 2.3 GB. A major computational cost of this project was acquiring the data and creating a local copy of Wikipedia to process. We would like to thank the Computation Science Center at CU Boulder for making available the resources for our research.

We would also like to thank Jim Martin and Martha Palmer for teaching us everything we know about Natural Language Processing.

8. REFERENCES

[1] R. L. Bangert-Drowns. The word processor as an instructional tool: A meta-analysis of word processing in writing instruction. *Review of Educational Research*, 63(1):69–93, 1993.

[2] A. Borthwick, J. Sterling, E. Agichtein, and R. Grishman. Exploiting diverse knowledge sources via maximum entropy in named entity recognition, 1998.

[3] S. Della Pietra, V. Della Pietra, and J. Lafferty. Inducing features of random fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(4):380–393, April 1997.

[4] P. W. Foltz. Supporting content-based feedback in on-line writing evaluation with lsa. *Interactive Learning Environments*, pages 111–127, August 2000.

[5] B. R. Gaines. An ounce of knowledge is worth a ton of data: quantitative studies of the trade-off between expertise and data based on statistically well-founded empirical induction. In *Proceedings of the sixth international workshop on Machine learning*, pages 156–159, San Francisco, CA, USA, 1989. Morgan Kaufmann Publishers Inc.

[6] J. Giles. Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901, December 2005.

[7] B. J. Grosz, S. Weinstein, and A. K. Joshi. Centering: a framework for modeling the local coherence of discourse. *Comput. Linguist.*, 21(2):203–225, June 1995.

[8] M. A. Hearst. Texttiling: A quantitative approach to discourse segmentation. Technical Report S2K-93-24, University of California, Berkeley, 1993.

[9] S. W. History. Wikipedia: Featured article criteria, May 2007.

[10] S. W. History. Wikipedia: Version 1.0 editorial team/index, May 2007.

[11] D. Jurafsky and J. H. Martin. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall, Upper Saddle River, N.J., 2000.

[12] N. Macdonald, L. Frase, P. Gingrich, and S. Keenan. The writer's workbench: Computer aids for text analysis. *Communications, IEEE Transactions on [legacy, pre - 1988]*, 30(1):105–110, 1982.

[13] D. Mehegan. Bias, sabotage haunt wikipedia's free world. *Boston Globe*, February 2006.

[14] R. Mihalcea. Using wikipedia for automaticword sense disambiguation. In *Proceedings of NAACL HLT 2007*, page 196.203. Association for Computational Linguistics, 2007.

[15] E. Miltsakaki and K. Kukich. Evaluation of text coherence for electronic essay scoring systems. *Nat. Lang. Eng.*, 10(1):25–55, March 2004.

[16] K. Nigam, J. Lafferty, and A. Mccallum. Using maximum entropy for text classification, 1999.

[17] J. Nocedal. Updating quasi-newton matrices with limited storage. *Mathematics of Computation*, 35(151):773–782, 1980.

[18] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[19] R. Pirsig. *Zen and the art of motorcycle maintenance :an inquiry into values*. Morrow, New York, 1974.

[20] S. P. Ponzetto. Creating a knowledge base from a collaboratively generated encyclopedia. In *Proceedings of NAACL HLT 2007*, pages 9–12. Association fo Computational Linguistics, 2007.

[21] S. P. Ponzetto and M. Strube. Creating a knowledge base from a collaboratively generated encyclopedia. In *Proceedings of NAACL HLT 2006*, pages 192–199. Association fo Computational Linguistics, 2006.

[22] J. C. Reynar and A. Ratnaparkhi. A maximum entropy approach to identifying sentence boundaries. In *Proceedings of the fifth conference on Applied natural language processing*, pages 16–19, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[23] L. Richardson. *Beautiful Soup Documentation*, April 2007.

[24] R. Rosenfeld. A maximum entropy approach to adaptive statistical language modeling, 1996.

[25] G. Weiss and F. Provost. The effect of class distribution on classifier learning, 2001.

[26] S. P. Witte and L. Faigley. Coherence, cohesion, and writing quality. *College Composition and Communication*, 32(2):189–204, 1981.

[27] L. Zhang. *Maximum Entropy Modeling Toolkit for Python and C++*, December 2004.