Identification of Effective Predictive Variables for Document Qualities

Kwong Bor Ng

Queens College, CUNY. Kissena Boulevard, Flushing, NY 11367. Email: kbng@qc.edu

Rong Tang, Sharon Small & Tomek Strzalkowski

SUNY Albany, Western Avenue, Albany, NY 12222. Email: {tangr, small, tomek}@albany.edu

Paul Kantor, Robert Rittman, Peng Song, Ying Sun & Nina Wacholder

School of Communication, Information and Library Studies, Rutgers, The State University of New Jersey, 4 Huntington St., New Brunswick, NJ 08903.
Email: {kantor, rritt, psong, ysun, nina}@eden.rutgers.edu

We analyzed textual properties of documents to identify predictive variables for various document qualities by means of statistical and linguistic methods. We have created a collection of 1000 documents, each document has been judged in terms of nine document qualities (accuracy, reliability, objectivity, depth, author/producer credibility, readability, verbosity and conciseness, grammatical correctness, one-sided or multi-view.) Employing statistical analyses, we considered a kind of linear combination, asking (1) if it was possible to combine textual features linearly to predict document qualities; (2) what textual features had good predictive power; (3) what textual features were minimally required for prediction with a detection rate much better than the false alarm rate. We present several promising results, indicating that with a few number of textual features, we can predict various document qualities much better than chance.

Document Relevance vs Document Quality

In information retrieval and filtering studies, the primary concern is the relevance of a document with respect to a topic). Other document properties, such as level of difficulty (i.e., intellectual access), reliability of the sources of information, authoritativeness of content, etc., usually are not considered by current systems because these properties usually have no direct relationship with topical relevance (with exception of course, e.g., see Naumann 2002).

We believe some of these properties are as important as relevance in their contribution to the usefulness of a document. In some analytical settings, it is of interest to estimate these properties in addition to the primary estimate of their relevance to the topic. We wanted to study these document properties, especially those properties that are usually associated with the quality of a document.

The quality of a document may be considered as consisting of two distinct dimensions: presentation and content. The presentation dimension relates to whether the information, observations and judgments contained in a document are presented in a particular manner, with a particular kind of textual structure. Different genre of documents may have different desirable styles of presentation. For example, for a news article, the desirable presentation manner may be clarity and comprehensiveness, and the presentation structure may be some kind of inverted triangle with the conclusion of the reported event summarized in the first paragraph. The content dimension relates to whether the information, observations and judgments contained in a document are reliable, objective and unbiased. Different document genres may also differ in desirable content. For example, for an advertisement, diversity of opinions or multiple viewpoints may not be desirable. In our study, we would like to investigate those quality aspects that make some documents more desirable to information professionals than others.

As a preliminary study, we focused on news media professionals as representative of information professionals. Two focus group sessions were conducted during March and April of 2002. The first session included seven participants who were news professionals; a majority were from a local newspaper, the Albany Times Union. The second session included four professional news editors. Both sessions took 90 minutes; the first was more of a free-form discussion and the second more of a task-oriented

meeting. As a result of the sessions, nine information quality criteria were generated. They are (see Appendix A for the definitions of these nine document qualities.):

- Accuracy,
- Source reliability,
- Objectivity,
- Depth,
- Author/Producer credibility
- Readability
- Verbose and conciseness
- Grammatically Correctness,
- One-Sided Multi-Views.

When a user expresses a preference with regard to some document qualities, an automatic means of estimating those qualities and classifying retrieved documents for presentation to the user is needed. This means that the estimation must be based on automatically computable characteristics of the document, whether statistical or linguistic.

We wanted to investigate how effective it would be for a machine to automatically estimate these qualities of a document. Employing statistical analyses, we considered a kind of linear combination, asking (1) Do some document qualities co-vary with certain textual features, such that an algorithm could be developed to estimate qualities based solely on textual features; (2) what textual features have good predictive power associated with certain qualities of a document, such as some parts of speech which could be automatically extracted from texts, or some linguistic patterns which could only be identified by human intelligence; (3) what textual features were minimally required for a prediction of document quality with a detection rate much better than the false alarm rate. We found that with a few textual features, we could predict various document qualities much better than chance. In the following section, we will present several promising results of our study.

Data Collection: Reliability and Validity

For this experiment we collected one thousand medium sized (100 words to 2500 words) news articles from the TREC collection (Voorhees, 2001), including articles from the LA Times, Wall Street Journal, Financial Times of London, and the Associated Press. We recruited two groups of news professionals and researchers to assign quality scores to the articles (using a Likert scale of 1 to 10). The techniques used to elicit these judgments from the subjects and the assessment of the reliability of the judgments is reported in Tang at al (2003)

Our analysis of the results of these focus groups produced 1000 quality vectors a quality vector consists of the nine quality variables, each with values equal to the average of two quality scores assigned by two judges of the collection.

By applying principal component analysis (Reyment & Joreskog, 1993) on these vectors we identified two major components (also reported in Tang *at al* 2003) that are consistent with our understanding of the nature of document quality. The horizontal component includes "credibility", "source reliability", "accuracy", "multi-view", "depth", and "objectivity", corresponding to the content dimension. The vertical component consists of "grammar", "readability", and "verbosity and conciseness", corresponding to the style dimension. Together they can explain about 58% of the variances of the nine quality variables (Figure 1).

Based on these results, we believe the reliability and validity of our data are high enough for us to conduct further analyses. In the following, whenever we refer to the score of a quality variable, we mean the average of the two scores assigned by two judges.

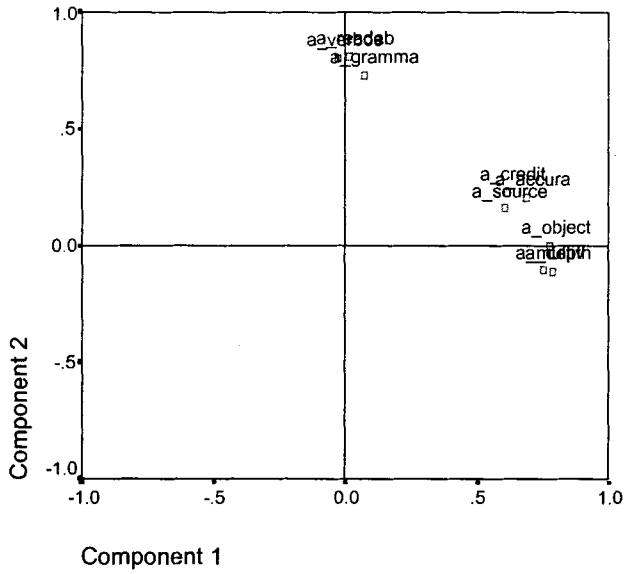

FIG. 1. Component plot of the nine quality variables using principal component analysis, in rotated space. (Rotation Method: Oblimin with Kaiser Normalization. Rotation converged in 5 iterations.)

Preliminary Data Analysis

Our objective is to find some textual features that are highly correlated with the quality variable for our collection of 1,000 judged documents. The textual features we considered at this initial stage were summarized in Table 1 (altogether 104 textual feature variables).

TABLE 1. Categories of textual features we used in the preliminary stage of data analysis.

Punctuation	Number of periods, question marks, exclamation marks, commas, dash, semicolons, colons, ellipsis, marks, parentheses, brackets, quotation forward slides, apostrophes, hyphens
Symbol	Number of dollar signs, percent signs, plus signs, > marks, ampersands
Length	Average length of words in characters, sentence in words, paragraph in words. Length of title, subtitle, leading paragraph, and document
Upper Case	Number of all upper case words, number of words with first character upper case
Quotation	Average quotation length
Key Terms	Number of word "say", "seem", and "expert"
Unique words	Number of unique words, number of unique words excluding stop words
POS	Number of token, proper noun, personal pronoun, possessive pronoun, determiner, preposition, verb in base form, verb in past tense, verb in past participle,verb in present participle, , verb in present tense, verb in ing form
Entities	Number of person, location, organization, and date.

We used existing grammatical part-of-speech (POS) taggers and information extraction tools developed for GATE (General Architecture for Text Engineering, see Cunningham et al., 2000) and AWB (Alembic Workbench, see Day et al, 1997), for POS extraction. The result seemed promising. For each of the nine quality variables, we could always find some highly correlated textual features. For example, see Table 2.

TABLE 2. High correlation between quality variables and textual features.

Quality Variable	**Textual Feature**	**Significance of correlation (2 tails)**
Accuracy	Personal Pronoun	0.0002
Reliability	Distinct organization	0.0048
Objectivity	Pronoun	0.0001
Depth	Document length	0.0000
Credibility	Date unit, e.g. day, week	0.0000
Readability	Closing parenthesis	0.0099
Verbose and conciseness	Subordinating preposition or conjunction	0.0003
Multi-view	Past form verb	0.0000
Grammatical correctness	Average length of paragraph in words	0.0016

Some of the high correlations are intuitively sensible (e.g., the high correlation between "depth" and document length), some are not (e.g., the high correlation between objectivity and number of pronouns). To pursue in this direction aggressively, we wanted to see if we could use multiple linear regression to estimate the values of each quality variable based on the textual features. In other words, we wanted to see if we could establish the following model:

$$\mathbf{Q_j} = \beta_0 + \sum \beta_i \mathbf{x_i}$$

where $\mathbf{Q_j}$ is the $\mathbf{j^{th}}$ quality variable and $\mathbf{x_i}$ is the $\mathbf{i^{th}}$ textual feature. The result was not very good. With the 104 textual features as independent variables, the total variances that could be explained were only 16.0% - 28.5% (Table 3).

TABLE 3. Goodness of Fit of Multiple Linear Regression.

Quality Variable	**R Square of Regression**
Accuracy	0.181
Reliability of Source	0.167
Objectivity	0.242
Depth	0.280
Credibility	0.237
Readability	0.262
Verbose and conciseness	0.210
Multi-view	0.285
Grammatical correctness	0.160

Using linear regression, the highest percentage of variance we can explain was just 28.5% (Multiple viewpoints). To understand what happened, we looked at the scatter plots of regression scores vs. actual scores of each of the 9 quality variables. They have a similar display. Figure 2 is representative. It is the scatter plot of the "regression scores" vs. "actual scores" of quality variable "depth".

As indicated in Figure 2, although there is a discernable trend, it would be quite difficult to use regression scores for actual score prediction. We speculated that if we divided the averaged (actual) scores into two ranges: high and low, with low = 0.5 – 5 and high = 5.5 – 10, and used the regression scores to predict high / low instead of the actual score, we might be able to do a much better job, as the number of the cases in the top right quarter (414 cases) is almost three times more than the bottom right quarter (106 cases), and the number of cases in the bottom left quarter (287 cases) is also significantly more than the top left quarter (193 cases).

We used linear discriminant analysis (Klecka, 1980) to test this idea. We randomly split our collection into two halves, one as training data and the other as testing data. In the training data set, we sought a linear combination of the frequencies of textual features (denoted by x_i in the following equation) as the basis for assigning cases into the high group and low group: $\mathbf{D_j} = \alpha_0 + \sum \alpha_i x_i$. Where $\mathbf{D_j}$ is the discriminant score of the quality variable **j**. Alpha was chosen in such a way that the ratio of between-group sum of squares to the within-group sum of squares would be at maximum. Then we applied the discriminant function to the testing data set to classify documents into high and low groups.

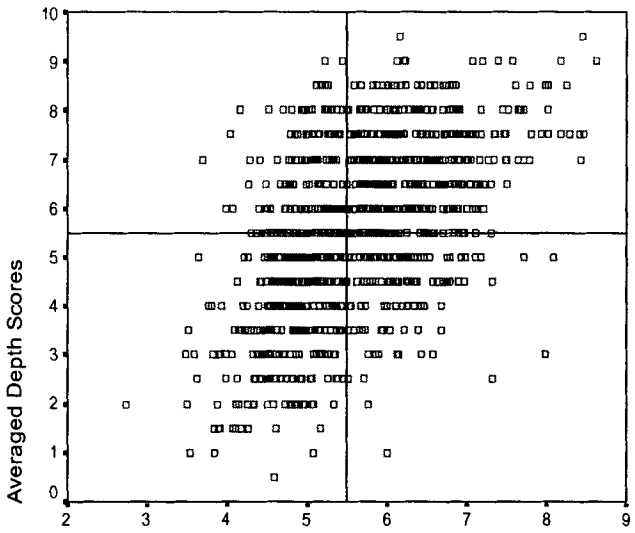

FIG. 2: Scatter plot of averaged depth scores vs. predicted scores calculated by multiple regression

The correct classification rate could be used as a measure of performance of this method. We repeated the experiment for each quality variable. The results were good, with correct classification rate better than chance, but not very impressive. Table 4 and Table 5 are the summaries of the two quality variables: depth and objectivity. The percentage in each cell is the correct classification rate of corresponding prediction.As we can see from Table 4 and Table 5, the power of the discriminant functions decreased more than 10% from the training data set to the testing data set, with only a little bit more than 60% of correct classification rate.

There are some inherent limitations associated with discriminant analysis. For example, it requires a multivariate Gaussian distribution which sometimes may not be guaranteed for the array of variables based on textual characteristics. Therefore, we supplemented discriminant analysis with logistic regression (Menard, 1995).

TABLE 4. Classification result of quality variable Depth. Overall, 74.5% of training cases can be correctly classified, 61.60% of testing cases correctly classified.

Depth			**Predicted Group Membership**	
			Low	High
Training Cases	Original	Low	67.7%	32.3%
		High	21.0%	79.0%
Testing Cases	Original	Low	54.4%	45.6%
		High	33.9%	66.1%

TABLE 5: Classification result of quality variable Objectivity. Overall, 75.5% of training cases correctly classified, 63.5% of testing cases correctly classified.

Objectivity			**Predicted Group Membership**	
			Low	High
Training Cases	Original	Low	58.7%	41.3%
		High	12.7%	87.3%
Testing Cases	Original	Low	45.5%	54.5%
		High	23.5%	76.5%

In our logistic regression analysis, we tried to establish the following model:

$$\log_e \frac{p_j}{1-p_j} = \alpha_0 + \sum \alpha_i x_i$$

where p_j is the probability of a document to have a high value of the quality variable j , x_i is the i^{th} textual feature, Alpha was chosen in such a way that the number of correction predictions would be at maximum. We ran logistic regression on all the quality variables, all gave us

correct classification rates similar to discriminant analysis (see our other paper, Tang *at all* 2003).

Human Understanding of Linguistic Meaning and Human Inconsistence

The predictive variables in the previous analyses were POS tags and other features identified by the GATE and AWB systems; this analysis did not take into consideration judges' perception of the relationship between text and document quality. In our data collection process, in addition to asking judges to assign quality scores, we had also asked judges to highlight and save portion of text as supportive evidence of their judgment of each of the nine qualities. In the body of the evidence, we noticed that they included many explicit declarations, implicit assertions, and references to the entities (people or organizations) that made the declarations / assertions. We therefore added three new categories of predictive variables (Table 6, see Appendix B for a comprehensive list.).

TABLE 6: Additional categories of textual feature

Explicit Declaration	Assertion made by subject. The subject is usually represented as a proper noun (personal name) or pronoun. The statement is distinguished by a verb such as "say", "declare", or "announce". The list of declaration verbs came from the saved evidence, and served as the basis of an expanded list created by using WordNet. Related variables are: Number of declarations, distance between a declaration verb and the verbal subject, number of cases that the verbal subject is in front of a declaration verb, number of cases that the verbal subject is after the declaration verb.
Other Assertion	This second list of variable is meant to represent "statements" that represent the asserters state of mind ; examples are "think", "believe", "decide", and "feel". The list of declaration verbs came from the evidence saved by judges, and served as the basis of an expanded list created with WordNet .
Entities other than human proper namesl	Four types of entities (organizations, political entities, locations and professional titels) were used to supplement the GATE entity lists. They were non-person entities making statements such as "the State Department said", "the president said", and "Beijing stated that" – extracted from the evidence saved by judges

One of the authors (RR) has developed a list of declarative verb based on the evidence. We called this R words (see Table 7). He also noticed that there were many proper and common nouns in the evidence. Based on this observation, he constructed a computed variable, which was the sum of the proper and common nouns, normalized by natural logarithm. To verify the importance of these new variables, we calculated the Pearson product moment correlation between each new variable and each document quality. We found that all the newly created variables are very highly correlated with the document qualities (Table 7).

TABLE 7: Pearson product moment correlation between new variables and document qualities.

New Variables	Depth	Reliability	Objectivity	Credibility	Readability	Verbose	Multi-view	Grammar
Natural log of sum of nouns	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**
Hitiqa organization	**.000**	**.000**	.241	**.018**	.330	.138	**.000**	.201
Hitiqa political	.237	.929	.711	.081	**.010**	.241	**.001**	.235
Hitiqa professional	**.003**	**.009**	.210	.288	**.003**	.759	**.002**	.868
Hitiqa country	**.000**	.056	**.044**	.459	.986	.114	**.000**	.566
Averaged declarative link distance[1]	**.040**	**.046**	**.001**	.537	**.000**	**.000**	**.006**	.717
Number of declarative verbs	**.000**	**.000**	**.000**	**.003**	**.000**	.802	**.000**	.005
Number of backward situation [2]	**.000**	**.000**	**.000**	**.021**	**.000**	.370	**.000**	**.007**
Accumulated link distance for backward situation	.416	.467	**.015**	.501	**.006**	**.000**	.281	.152
Number of forward situation [3]	**.000**	**.000**	**.000**	**.001**	**.000**	.921	**.000**	**.008**
Accumulated link distance for forward situation	**.007**	.946	.790	.418	.355	**.006**	.123	.212
Number of declarative verb (R words list)	**.000**	**.000**	**.000**	**.001**	**.000**	.329	**.000**	**.011**
R words list minus [4]	**.000**	**.000**	**.000**	**.001**	**.000**	.377	**.000**	**.005**
R words list miplus [5]	**.000**	.930	.183	.996	.469	**.000**	**.001**	.123

[1] Averaged distance (in number of words) between declarative verbs and the subject that made the declaration.

[2] The subject that made the declaration was found when searched backward from the declarative verb.

[S] The subject that made the declaration was found when searched forward from the declarative verb.

[4] Count of hyponyms of words in R list found by wordnet and are included in R list

[5] Count of hyponyms of declarative verbs from R words list, minus number of declarative verb from R words list

In Table 7, the left most column is the new variable. The variables hitiqa organization, hitiqa political, hitiqa professional and hitiqa country are based on the corresponding terms in the evidence saved by judges. We used those terms as seeds in WordNet (Fellbaum 1998) to find out all their children. The variables are the frequencies of the parents and children in documents.

In the body of the table, each cell records the level of significance (two tailed test) of the correlation. Those with significance level less than 0.05 are in bold face. From this result, we expected that, if we included these variables into our list of predictive variables, there should be a dramatic improvement in correct classification rate. We were wrong.

With the addition of these new predictive variables, we had a slight improvement in our prediction using discriminant analysis as well as logistic regression across all qualities, but the improvement was too small to be significant. For example, for the quality "depth" (Table 8), the correct classification in the training data set (using discriminant analysis) was improved by only 2.6% (from 74.5% to 77.1%) and the correct classification in the testing data set was improved by 1.3% (from 61.6% to 62.9%).

TABLE 8. Classification result of quality variable Depth. Overall, 77.1% of training cases can be correctly classified, 62.9% of testing cases correctly classified.

Depth			**Predicted Group Membership**	
			Low	High
Training Cases	Original	Low	68.7%	31.3%
		High	17.3%	81.7%
Testing Cases	Original	Low	54.4%	45.6%
		High	31.6%	68.4%

For the other quality variables, the correct classification (using discriminant analysis) rates of training and testing were similar to the previous test of "depth" with the new variables.

Since the dimension of human understanding had entered our picture, we continued to try to explore in the direction of the role played by human judgment. There is always a subjective dimension to human judgment (that partially explains the inter-judge inconsistency) and, different subjective concerns or a subjective background, may result in disagreement. When a machine looked at the averaged score, it would not consider such differences. Scores with a large variance would be treated as a small variance if they had the same mean. We suspected that, instead of asking a machine to learn only the average behavior of two judges (the averaged scores), we should ask the machine to give more attention to those judgments that agreed with each other and less attention to those judgments that disagreed. In addition, if a machine could not predict correctly the membership (high or low in a particular quality) of a document about which two judges disagreed, it should not be counted as a total failure.

Therefore, we added a weight factor to our machine learning and testing. Each document had nine weights (corresponding to nine quality variables), each weight was equal to $\frac{1}{(q_{i1}-q_{i2})^2+1}$, where q_{i1} was the score of the i^{th} quality variable assigned by one judge and q_{i2} was the score of the same quality variable assigned by the other judge.

If two judges agreed with each other, q_{i1} would be close to q_{i2}, and the weight would be approximately equal to 1; the more two judges disagreed, the lower the weight. If two judges totally disagreed (i.e., one assigned the value 1 and the other assigned the value 10), the weight would be 0.012, and the case would become negligible.

For the document quality "depth", using the weight mechanism with discriminant analysis, correct classification was improved by 82.6% - 77.1% = 5.5% in the training data set, but when we applied the discriminant function to the testing data set, the performance only increased by 63.3% - 62.9% = 0.4%.

In evaluating this result, we wondered if we might not have enough cases in the training data set. In corpus analysis, using 500 cases for training might not be enough to predict another 500 cases. To compensate for the possibility that we might not have a large enough corpus, we used 800 cases for training and 200 cases for testing. The results were better. For example, for the quality "depth", the correct classification rate was improved by 79.7% - 77.1% = 2.6% in the testing data set and by 67.6% - 62.9% = 4.7% in the testing data set, as shown in Table 9. As we see from Table 9, the model performed much better in identifying high-depth documents than low-depth documents.

TABLE 9. Classification result of quality variable "Depth". Overall, 79.7% of training cases can be correctly classified, 67.6% of testing cases correctly classified.

Depth			**Predicted Group Membership**	
			Low	High
Training Cases	Original	Low	64.5%	35.5%
		High	11.9%	88.1%
Testing Cases	Original	Low	51.0%	49.0%
		High	22.6%	75.4%

Predictive Variable Reduction

In our study, discriminant analysis was used as an exploratory tool. In order to arrive at a good model, all the potentially useful variables were included in the data set. It was not known in advance which of these variables were important for high / low distinction and which were, more or less, extraneous. One of the desired products of the analysis was the identification of the "good" predictor variables. Therefore, we decided to apply a stepwise variable selection algorithm to see if we could reduce the number of predictive variables.

In the stepwise discriminant analysis (Huberty 1994), the first variable included in the analysis had the largest acceptable value for the selection criterion. After the first variable was entered, the value of the criterion was re-evaluated for all variables not in the model, and the variable with the largest acceptable criterion value was entered next. At this point, the variable entered first was re-evaluated to determine whether it met the removal criterion. If it did, it was removed from the model. The next step was to examine the variables not in the equation for entry, followed by examination of the variables in the equation for removal. Variables were removed until none remained that met the removal criterion. Variable selection terminated when no more variables met entry or removal criteria.

Using the stepwise method, we could reduce the number of predictive variables from more than a hundred to just a few, and the predictive power of the models was as good as using all variables. For example, in predicting document depth, stepwise discriminant analysis reduced the number of predictive variables to just four :

- log of sum of nouns
- date unit, e.g., day, week
- determiner
- number of all upper case words.

Using the discriminant equation, overall, 70.9% of training cases could be correctly classified, 75.6% of testing cases correctly classified! We were somewhat surprised that the classification rate in the testing stage was better than the training stage.

The ROC (Receiver Operating Characteristic) curve is a good summary of the probability distribution because it will give us not only the discriminant power, but also a basis for selecting specific cutoff points (Egan, 1975). If we consider classifying high-depth documents correctly as "detection" and classifying low-depth documents incorrectly as high-depth documents as "false alarm", we can plot two ROC curves by sorting discriminant scores numerically. To understand this, we plotted the ROC curve (Figure 3). Every point along the curves represents a possible cutoff point to discriminate between high-depth and low-depth documents. Each point represents a detection rate and a false alarm rate. The associated detection rate of a point is the ratio of the number of high-depth documents that would be correctly classified using that point as a cutoff to the total number of high-depth documents. The associated false alarm rate of a point is the ratio of the number of low-depth documents that would be incorrectly classified as high-depth using that point as a cutoff to the total number of low-depth documents.

As illustrated in Figure 3, the two curves are very similar. The training curve is above the testing curve in some ranges, and the testing curve above the training curve in other ranges. Because the two curves are so close, it is not surprising that there are points in the testing curve with higher detection rates and lower false alarm rates than the training curve.

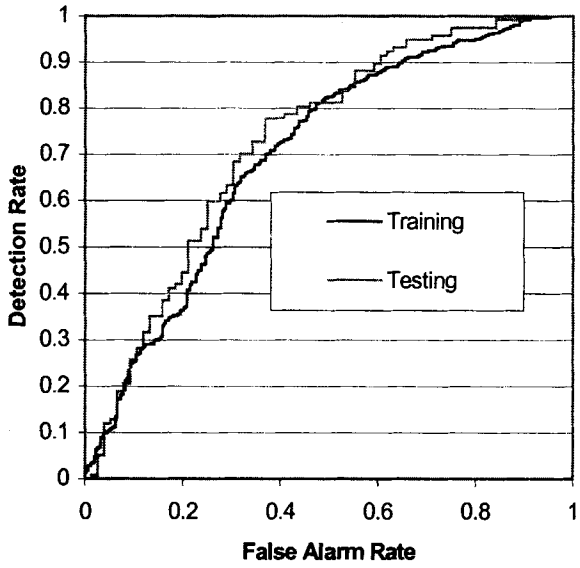

FIG. 3. ROC curves of training and testing (quality variable: depth) using stepwise discriminant function.

We observed similar results for other quality aspects. For example, using stepwise discriminant analysis for "objectivity" reduced the predictive variables from more than a hundred to only eight :

- average length of paragraph in words
- number of forward situations
- distinct organizations
- possessive pronouns
- comparative adjectives
- plural proper nouns
- distinct persons
- length of leading paragraph.

Overall, the discriminant equation correctly classified 73.3% of training cases, and 71.3% of testing cases.

Conclusion

By reducing more than 100 machine-computable characteristics of text to a manageable set, we identified two guiding principles. One principle is to select features that "make sense" from a linguistic point of view. The other is to let a statistical package select those variables on a step-wise basis that have the greatest power in predicting the target characteristic. Extensive analyses using features identified by two different natural language tagging packages (AWB and GATE) showed that these approaches could complement each other. As seen in the detailed analytical tables, it is possible to construct, after the fact, a plausibility argument for a statistically selected feature.

In our exploration, we have shown that linear combinations of just a few textual features can predict some qualities of documents, with detection rates higher than the false alarm rate. We found that using common textual features alone might not be enough to identify good predictors. Adding a set of new computed variables (based on our understanding of the evidence), not only improved the correct prediction rate, but the stepwise method preserved many of these new variables in the final set of good predictors of various qualities. Our next step will seek to identify additional computed variables based on a better understanding of the evidence.

ACKNOWLEDGEMENTS

This paper is based on work supported by the Advanced Research and Development Activity (ARDA)'s Advanced Question Answering for Intelligence (AQUAINT) Program. The views expressed in this article are those of the authors, and do not necessarily represent the views of the sponsoring agency.

We gratefully acknowledge the contribution of all the participants of the quality experiments at both Albany and Rutgers. We also wish to thank Ting Liu, Nobuyuki Shimizu, Uzo Enyinna, and Tom Palen for their contribution to the project.

Appendix A: Definition of the Nine Document Qualities

Accuracy: The extent to which information is precise and free from known errors.

Source Reliability: The extent to which you believe that the indicated sources in the text (e.g., interviewees, eye-witnesses, etc.) provide truthful account of the story.

Objectivity: The extent to which the document includes facts without distortion by personal or organizational biases.

Depth: The extent to which the coverage and analysis of information is detailed.

Author/Producer Credibility: The extent to which you believe that the author of the writing is trustworthy.

Readability: The extent to which information is presented with clarity and is easily understood.

Verbose → Conciseness: The extent to which information is well-structured and compactly represented.

Grammatical Correctness: The extent to which the text is free from syntactic problems.

One-sided → Multi-views: The extent to which information reported contains a variety of data sources and view points.

REFERENCES

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., & Wilks, Y. (2000). Experience of using GATE for NLP R&D. In Workshop on Using Toolsets and Architectures To Build NLP Systems at COLING-2000, Luxembourg.

Day, D., Aberdeen, J., Hirschman, L., Kozierok, R., Robinson, P. & Vilain, M. (1997) Mixed-initiative development of language processing systems. In Fifth Conference on Applied Natural Language Processing, Association for Computational Linguistics. Retrieved December 12, 2002 from, http://www.mitre.org/technology/alembic-workbench/ANLP97-bigger.html

Egan, J.P. (1975). Signal detection theory and ROC analysis. New York: Academic Press.

Fellbaum,, C. (ed). (1998). WordNet: An Electronic Lexical Database. The MIT Press.

Huberty, C. J. (1994). Applied Discriminant Analysis. Wiley-Interscience.

Klecka, William, R. (1980) Discriminant analysis. Sage University Paper series on quantitative applications in the social sciences, series no. 07-019. Thousand Oaks, CA: Sage.

Menard, Scott (1995) Applied logistic regression analysis. Sage University Paper series on quantitative applications in the social sciences, series no. 07-106. Thousand Oaks, CA: Sage.

Naumann, F. (2002). Quality-driven Query Answering for Integrated Information Systems. Berlin: Springer-Verlag.

Reyment, R. & Joreskog, K.G. (1993) Applied factor analysis in the natural science. Cambridge Univer-sity Press.

Rong, T. Ng, K.B., Strzalkowski, T. & Kantor, P. (2003) Toward Machine Understanding of Information Quality. In ASIS 2003 Annual Meeting Proceedings (i.e., same volume).

Voorhees, E. (2001). Overview of TREC 2001. In E. Voorhees (ed.) NIST Special Publication 500-250: The Tenth Text REtrieval Conference, pp. 1 – 15. Washington, D.C.