Document Features Predicting Assessor Disagreement

Praveen Chandar  William Webber  Ben Carterette
Department of Computer & Information Sciences  College of Information Studies  Department of Computer & Information Sciences
University of Delaware  University of Maryland  University of Delaware
Delaware, USA  Maryland, USA  Delaware, USA
pcr@cis.udel.edu  wew@umd.edu  carteret@cis.udel.edu

ABSTRACT
The notion of relevance differs between assessors, thus giving rise to assessor disagreement. Although assessor disagreement has been frequently observed, the factors leading to disagreement are still an open problem. In this paper we study the relationship between assessor disagreement and various topic-independent factors such as readability and cohesiveness. We build a logistic model using reading level and other simple document features to predict assessor disagreement and rank documents by decreasing probability of disagreement. We compare the predictive power of these document-level features with that of a meta-search feature that aggregates a document's ranking across multiple retrieval runs. Our features are shown to be on a par with the meta-search feature, without requiring a large and diverse set of retrieval runs to calculate. Surprisingly, however, we find that the reading level features are negatively correlated with disagreement, suggesting that they are detecting some other aspect of document content.

Categories and Subject Descriptors
H.3.4 [Information Storage and Retrieval]: Systems and software—performance evaluation.

Keywords
Retrieval experiment, evaluation

General Terms
Measurement, performance, experimentation

1. INTRODUCTION
Human assessors are used in information retrieval evaluation to judge the relevance of a document for a given topic. Assessors frequently disagree on the relevance of a document to a topic, however. A study by [7] found that the probability that a second assessor would agree with a first assessor’s judgment that a document was relevant was only two in three. A survey of such studies done by [2] found similar results as well. While [7] found that assessor disagreement had limited effect on the comparative evaluation of systems, it does have a major impact upon the evaluation of their absolute effectiveness. Moreover, a simulation study by [4] suggests that the effect on comparative evaluation depends upon the nature of disagreement, and that an overly liberal (or careless) assessor introduces considerable noise even to the comparison of retrieval systems.

While assessor disagreement has been frequently observed, and its effect on retrieval evaluation somewhat studied, less work has been done on the factors that lead to assessor disagreement. [9] observes that there is great variability in disagreement between different assessor pairs and on different topics. Regarding assessor-level effects, [8] find that assessor training has little effect on reliability (legally trained assessors no more than untrained assessors on e-discovery tasks). Regarding topic-level effects, [11] find that more detailed assessor instructions do not seem to increase disagreement.

In addition to assessor-level and topic-level effects on assessor disagreement, there may be document-level effects: some documents may be more likely to provoke assessor disagreement than others. [10] have begun work in this direction, using metarank information across multiple runs to predict disagreement. If one assessor finds a document relevant, but it is generally lowly ranked by retrieval systems, then a second assessor is likely to disagree with the original assessor, and conversely with originally irrelevant but highly-ranked documents.

In the current paper, we investigate the relation between assessor disagreement and various topic-independent document features. One set of such features are various metrics of the reading level or reading difficulty of a document. Our hypothesis is that documents that are more difficult to read will provoke higher levels of assessor disagreement. We also consider document length (hypothesizing that longer documents will provoke more disagreement) and document coherence (hypothesizing that less coherent documents will provoke more disagreement). Finally, we extend the metarank method of [10] by considering not only average rank across different retrieval systems, but also the variability in the ranking—using disagreement between retrieval systems as a predictor of disagreement between human assessors.

If reliable document-level predictors of assessor disagreement can be found, then they can be used to efficiently direct multiple assessments towards those documents most likely to provoke assessor disagreement. We consider this as a ranking problem, in which documents must be ranked by decreasing probability of assessor disagreement, examining the case in which this ranking must be made without any initial relevance assessment having been performed. Our experimental results indicate that document-level features give a significant improvement over random choice in predicting assessor disagreement. Moreover, where initial relevance

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
SIGIR '13, July 28–August 1, 2013, Dublin, Ireland.
Copyright 2013 ACM 978-1-4503-2034-4/13/07...
$15.00.
assessments are not available, document-level features predict as-
sever disagreement as strongly as meta-rank features, without re-
quiring a large and diverse set of retrieval runs to calculate.

One surprise of the study is that while reading level features are
predictive of assessor disagreement, the correlation is the opposite
of that posited in our hypothesis: documents scored as easier to
read are more, not less, likely to provoke assessor disagreement
than those scored as difficult to read. This suggests that reading
level features are themselves correlated with some other aspect of
document construction or content, which if more directly identified
could lead to even stronger predictors of assessor disagreement; a
question which is left to future work.

The remainder of the paper is structured as follows. A descrip-
tion of our logistic regression model along with all the document-
level features is given in Section 2.2. Section 3 describes our ex-
periments along with the dataset used in this work, and a detailed
analysis of our results is given in Section 4. Section 5 summarizes
our findings and sketches future work.

2. ASSESSOR DISAGREEMENT

Our approach to the problem of predicting assessor disagreement
consists of two main components: identifying features, and de-
veloping a modeling technique.

2.1 Logistic regression

We predict the probability that a document will attract diver-
gent binary relevance assessments from two or more assessors \((D, s)\),
based upon various document level features \(s = \langle s_i \rangle\), as \(p(D = 1|s)\). As we are predicting a probability, it is natural to apply a
logistic regression to this problem:

\[
p(D = 1|s_i) = \frac{e^{\beta_0 \cdot \text{aveWordLength}}}{1 + e^{\beta_0 \cdot \text{aveWordLength}}}
\]

where \(s_i\) is the score for feature \(i\), and the probability \(p\) is the pre-
dicted value. The fitted value \(\beta_0\) in Equation 1 is the intercept,
which gives the log-odds of disagreement when the score is 0, while
\(\beta_i\) is the score coefficient for feature \(i\), which gives the change or
“slope” in log odds of disagreement for every one point increase in
the given feature scores. The slope gives the strength of relation-
ship between feature scores and probability of disagreement, while
intercept the shifts the regression curve up or down the score axis.

A model can be built for each topic individually, or a univer-
sal model can be built using all queries in our dataset. The degree
to which a universal model is a good approximation for per-topic
models depends upon the strength of per-topic factors in influen-
cing disagreement. The closer the universal model is to the per-
topic models, the more likely is that a generalized model can be
built, that is able to predict assessor disagreement on new collec-
tions based only the feature scores.

2.2 Document Features

In this section, we discuss in detail the various predictors that
we use in Equation 1 to estimate assessor disagreement. The lo-
gistic model described in Section 2.1 relies heavily on the feature
scores and identifying good predictors of disagreement is critical.
We use a combination of simple document characteristic features
and reading level features to estimate disagreement.

2.2.1 Simple Document Features

The simple document quality features are described below:

- **docLength** Total number of terms in a document is a simple
feature that estimates the amount of information available in the
document.

- **aveWordLength** Average word length (number of characters)
is a very simple estimate of readability of a document.

- **Entropy** An estimate of document cohesiveness can be ob-
tained using the entropy of the document \([3]\). Document en-
tropy is computed over the words in the document as follows:

\[
E(D) = - \sum_{w \in D} P(w) \log(P(w))
\]

where \(P(w)\) can be estimated by the ratio of frequency of the
word to the total number of words in the document. Lower
entropy reflects a document that is focused on a single topic,
while higher entropy indicates a more diffuse document.

2.2.2 Reading Level Features

We employ a number of standard metrics of reading level, based
upon simple textual statistics. More complicated statistical and lan-
guage model approaches are left for future work \([5]\).

- **FleschIndex and Kincaid** are designed to capture the com-
prehension level of a passage. The two measures use word
and sentence length with different weighting factors. FleschIn-
dex is a test of reading ease with higher scores indicating text
that is easier to read. Kincaid is a grade score that is nega-
tively correlated to FleschIndex. A generic formula for both
metrics is given below:

\[
a \cdot \frac{\text{words}}{\text{sentences}} + b \cdot \frac{\text{syllables}}{\text{words}} + c
\]

where the values of \(a, b, c\) are as follows: FleschIndex \((a = -1.01, b = -84.6, c = 206.83)\) and Kincaid \((a =
0.39, b = 11.8, c = -15.59)\).

- **FogIndex** relies on average sentence length and the percent-
age of complex words for each passage of 100 words. Words
with three or more syllables are identified as complex words.

\[
0.4 \left( \frac{\text{words}}{\text{sentences}} + 100 \frac{\text{complexWords}}{\text{words}} \right)
\]

- **SMOG** (Simple Measure of Gobbledygook) was designed as an
easier and more accurate substitute to FogIndex, and is more
prevalent in the medical domain. It relies on two fac-
tors: the number of polysyllables (words with 3 or more syl-
lables) and the number of sentences.

\[
1.043 \sqrt{\text{numOfPolysyllables} \times \frac{30}{\text{sentences}}} + 3.120
\]

- **Lex** is a simple measure of readability computed by adding
average sentence length and number of long words. Words
with 6 or more letters are considered as long words.

\[
\frac{\text{words}}{\text{sentences}} + \frac{(\text{longwords} \times 100)}{\text{words}}
\]

- **ARI** (Automated Readability Index) is computed by com-
bining the ratio of the number of characters per word and
number of words per sentence. ARI relies on the number of
characters per word instead of syllables per word.

\[
4.71 \frac{\text{characters}}{\text{words}} + 0.5 \frac{\text{words}}{\text{sentences}} - 21.43
\]
method used in the disagreement ranking stage is described below.

development of a universal model for ranking by performing per-topic regressions (Section 4.1), then investigate the usefulness of these features as predictors of disagreement by building and testing universal (cross-validated) models (Section 4.2).

4. RESULTS AND ANALYSIS

We first analyze the relationship between individual features and assessor disagreement by performing per-topic regressions (Section 4.1), then investigate the usefulness of these features as predictors of disagreement by building and testing universal (cross-validated) models (Section 4.2).

4.1 Feature Analysis

We test our hypotheses that: (1) documents with higher comprehension difficulty, (2) longer documents, and (3) documents that are less focused on a topic (less cohesive), are more likely to be disagreed upon. For each feature, we build a logistic regression model on each topic with that feature as the single predictor, and observe the coefficients that the feature achieves across the 48 topics (the $\beta$ values in Equation 1). We calculate the average coefficient, and perform a two-sided, one sample t-test to test whether this coefficient differs significantly from zero across the 48 topics.

Table 4.1 reports our results. The metarank features are all highly significant. Entropy is also a significant positive predictor. In so far as entropy measures topic diffuseness, this confirms our hypothesis that more diffuse documents provoke higher levels of disagreement. Many of the reading level predictors also prove significantly correlated with disagreement. Surprisingly, however, the correlation is in the opposite direction from the hypothesis. Documents that get lower reading level scores, and therefore are marked as being easier to read, in fact provoke higher levels of assessor disagreement. (Recall that FleschIndex is the only reading level feature where higher scores mean easier comprehension.)

4.2 Modeling Disagreement

Next, we investigate how useful our method is at predicting assessor disagreement, using a universal (cross-validated) model to rank the documents of each topic by decreasing probability of assessor disagreement. Table 4.2 summarizes performances for average precision and precision at various cutoffs. We add as a baseline the expected precision achieved by a random sorting of the documents, which is just the macroaveraged proportion of disagreed documents per topic. A universal model that combines all our
features (denoted by “All Combined”) and a model that uses the metarank features significantly improves over random ordering under all measures. All the other features give a significant improvement over random order for MAP only, suggesting that top-of-ranking performance is mediocre. Entropy does best, as in Table 4.1, whereas the combined reading levels, despite being significant correlated with disagreement give very little benefit in terms of predicting disagreement under a universal model.

5. CONCLUSION

We started this paper with three hypotheses, namely that the documents that assessors are more likely to disagree on are: (1) documents with higher comprehension difficulty; (2) longer documents; and (3) documents that are less cohesive. At least in so far as these three conditions are captured by the measures we have used, our results have been mixed. The correlation between entropy and disagreement confirms the third hypothesis, and provides a weakly useful practical predictor of disagreement. The relationship between document length and disagreement (our second hypothesis), if it exists, is too weak for our experiments to detect as significant. Most surprisingly of all, our first hypothesis, that difficult documents would provoke more disagreement, has not only failed to be confirmed, but in fact the reverse has been observed: it is easier documents that provoke the most disagreement.

As it seems intuitively hard to believe that it is in fact easily-comprehended documents that assessors disagree the most about, a more likely interpretation of our results is that the reading level measures are picking up some other aspect of document content, syntax, or representation that tends to provoke disagreement in assessors. An informal examination of disagreed-upon documents that attracted easy reading level scores, for instance, suggests that a disproportionate number of them are transcripts of spoken text—presidential debates, speeches, interviews, and the like. These tend to have short sentences, but diffuse topics, and may be difficult to read quickly. Further work is to determine whether there are other text metrics that can more directly and accurately target the aspects of a document that predict assessor disagreement.

Acknowledgments: This material is based in part upon work supported by the National Science Foundation under Grant No. 1065250. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

6. REFERENCES


<table>
<thead>
<tr>
<th>Predictor</th>
<th>p-value</th>
<th>β_i</th>
</tr>
</thead>
<tbody>
<tr>
<td>FleschIndex</td>
<td>0.108</td>
<td>139.4</td>
</tr>
<tr>
<td>ColemanLau</td>
<td>0.163</td>
<td>-164.4</td>
</tr>
<tr>
<td>SMOGGrading</td>
<td>0.077</td>
<td>-166.4</td>
</tr>
<tr>
<td>Lix</td>
<td>0.012</td>
<td>-241.7</td>
</tr>
<tr>
<td>Kincaid</td>
<td>0.022</td>
<td>-133.3</td>
</tr>
<tr>
<td>ARI</td>
<td>0.006</td>
<td>-156.0</td>
</tr>
<tr>
<td>FogIndex</td>
<td>0.018</td>
<td>-159.2</td>
</tr>
<tr>
<td>docLength</td>
<td>0.052</td>
<td>51.2</td>
</tr>
<tr>
<td>aveWordLength</td>
<td>0.225</td>
<td>-374.7</td>
</tr>
<tr>
<td>Entropy</td>
<td>&lt; 0.001</td>
<td>832.1</td>
</tr>
<tr>
<td>metaAPSum</td>
<td>&lt; 0.001</td>
<td>159.7</td>
</tr>
<tr>
<td>metaAPsiDev</td>
<td>&lt; 0.001</td>
<td>206.8</td>
</tr>
<tr>
<td>metaAPMax</td>
<td>&lt; 0.001</td>
<td>321.2</td>
</tr>
</tbody>
</table>

Table 1: Results of significance test using two-sided one sample t-test with p-values and mean co-efficient scores across all 48 topics.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>P@5</th>
<th>P@10</th>
<th>P@20</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>random</td>
<td>0.216</td>
<td>0.216</td>
<td>0.216</td>
<td>0.216</td>
</tr>
<tr>
<td>metaAP</td>
<td>0.317*</td>
<td>0.350*</td>
<td>0.357*</td>
<td>0.372*</td>
</tr>
<tr>
<td>docLength</td>
<td>0.229</td>
<td>0.229</td>
<td>0.235</td>
<td>0.255*</td>
</tr>
<tr>
<td>Entropy</td>
<td>0.258</td>
<td>0.254</td>
<td>0.241</td>
<td>0.261*</td>
</tr>
<tr>
<td>aveWordLength</td>
<td>0.200</td>
<td>0.190</td>
<td>0.215</td>
<td>0.240*</td>
</tr>
<tr>
<td>ReadingLevel</td>
<td>0.246</td>
<td>0.252</td>
<td>0.229</td>
<td>0.239*</td>
</tr>
<tr>
<td>All Combined</td>
<td>0.321*</td>
<td>0.329*</td>
<td>0.341*</td>
<td>0.362*</td>
</tr>
</tbody>
</table>

Table 2: Performance Comparison at various ranks with significant improvement over expected random scores indicated by * (paired t-test). The results are based on 5-fold cross validation across 48 topics.