Automatic Assessment of Coverage Quality in Intelligence Reports

Samuel Brody
School of Communication
and Information
Rutgers University
sdbrody@gmail.com

Paul Kantor
School of Communication
and Information
Rutgers University
paul.kantor@rutgers.edu

Abstract

Common approaches to assessing document quality look at shallow aspects, such as grammar and vocabulary. For many real-world applications, deeper notions of quality are needed. This work represents a first step in a project aimed at developing computational methods for deep assessment of quality in the domain of intelligence reports. We present an automated system for ranking intelligence reports with regard to coverage of relevant material. The system employs methodologies from the field of automatic summarization, and achieves performance on a par with human judges, even in the absence of the underlying information sources.

1 Introduction

Distinguishing between high- and low-quality documents is an important skill for humans, and a challenging task for machines. The majority of previous research on the subject has focused on low-level measures of quality, such as spelling, vocabulary and grammar. However, in many real-world situations, it is necessary to employ deeper criteria, which look at the content of the document and the structure of argumentation. One example where such criteria are essential is decision-making in the intelligence community. This is also a domain where computational methods can play an important role. In a typical situation, an intelligence officer faced with an important decision receives reports from a team of analysts on a specific topic of interest. Each decision may involve several areas of interest, resulting in several collections of reports. Additionally, the officer may be engaged in many decision processes within a small window of time. Given the nature of the task, it is vital that the limited time be used effectively, i.e., that the highest-quality information be handled first. Our project aims to provide a system that will assist intelligence officers in the decision making process by quickly and accurately ranking reports according to the most important criteria for the task.

In this paper, as a first step in the project, we focus on content-related criteria. In particular, we chose to start with the aspect of "coverage". Coverage is perhaps the most important element in a time-sensitive scenario, where an intelligence officer may need to choose among several reports while ensuring no relevant and important topics are overlooked.

2 Related Work

Much of the work on automatic assessment of document quality has focused on student essays (e.g., Larkey 1998; Shermis and Burstein 2002; Burstein et al. 2004), for the purpose of grading or assisting the writers (e.g., ESL students). This research looks primarily at issues of grammar, lexical selection, etc. For the purpose of judging the quality of intelligence reports, these aspects are relatively peripheral, and relevant mostly through their effect on the overall readability of the document. The criteria judged most important for determining the quality of an intelligence report (see Sec. 2.1) are more complex and deal with a deeper level of representation.

In this work, we chose to start with criteria related to content choice. For this task,

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 491–495,
Portland, Oregon, June 19-24, 2011.

we propose that the most closely related prior research is that on automatic summarization, specifically multi-document extractive summarization. Extractive summarization works along the following lines (Goldstein et al., 2000): (1) analyze the input document(s) for important themes; (2) select the best sentences to include in the summary, taking into account the summarization aspects (coverage, relevance, redundancy) and generation aspects (grammaticality, sentence flow, etc.). Since we are interested in content choice, we focus on the summarization aspects, starting with coverage. Effective ways of representing content and ensuring coverage are the subject of ongoing research in the field (e.g., Gillick et al. 2009, Haghighi and Vanderwende 2009). In our work, we draw on elements from this research. However, they must be adapted to our task of quality assessment and must take into account the specific characteristics of our domain of intelligence reports. More detail is provided in Sec. 3.1.

2.1 The ARDA Challenge Workshop

Given the nature of our domain, real-world data and gold standard evaluations are difficult to obtain. We were fortunate to gain access to the reports and evaluations from the ARDA workshop (Morse et al., 2004), which was conducted by NIST in 2004. The workshop was designed to demonstrate the feasibility of assessing the effectiveness of information retrieval systems. During the workshop, seven intelligence analysts were each asked to use one of several IR systems to obtain information about eight different scenarios and write a report about each. This resulted in 56 individual reports.

The same seven analysts were then asked to judge each of the 56 reports (including their own) on several criteria on a scale of 0 (worst) to 5 (best). These criteria, listed in Table 1, were chosen by the researchers as desirable in a "high-quality" intelligence report. From an NLP perspective they can be divided into three broad categories: content selection, structure, and readability. The written reports, along with their associated human quality judgments, form the dataset used in our experiments. As mentioned, this work focuses on coverage. When assessing coverage, it is only meaningful to compare reports on the same scenario. Therefore, we regard our dataset as 8 collections (Scenario A to Scenario H), each containing 7 reports.

Content	
COVER	covers the material relevant to the query
NO-IRR	avoids irrelevant material
NO-RED	avoids redundancy
Structure	
ORG	organized presentation of material
Readability	
CLEAR	clear and easy to read and understand

Table 1: Quality criteria used in the ARDA workshop, divided into broad categories.

3 Experiments

3.1 Methodology

In the ARDA workshop, the analysts were tasked to extract and present the information which was relevant to the query subject. This can be viewed as a summarization task. In fact, a high quality report shares many of the characteristics of a good document summary. In particular, it seeks to cover as much of the important information as possible, while avoiding redundancy and irrelevant information.

When seeking to assess these qualities, we can treat the analysts' reports as output from (human) summarization systems, and employ methods from automatic summarization to evaluate how well they did.

One challenge to our analysis is that we do not have access to the information sources used by the analysts. This limitation is inherent to the domain, and will necessarily impact the assessment of coverage, since we have no means of determining whether an analyst has included all the relevant information to which she, in particular, had access. We can only assess coverage with respect to what was included in the other analysts' reports. For our task, however, this is sufficient, since our purpose is to identify, for the person who must choose among them, the report which is most comprehensive in its coverage, or indicate a subset of reports which cover all topics discussed in the collection as a whole[1].

[1]The absence of the sources also means the system is only able to compare reports on the same subject, as opposed to humans, who might rank the coverage quality

As a first step in modeling relevant concepts we employ a word-gram representation, and use frequency as a measure of relevance. Examination of high-quality human summaries has shown that frequency is an important factor (Nenkova et al., 2006), and word-gram representations are employed in many summarization systems (e.g., Radev et al. 2004, Gillick and Favre 2009). Following Gillick and Favre (2009), we use a bigram representation of concepts[2]. For each document collection D, we calculate the average prevalence of every bigram concept in the collection:

$$prev_D(c) = \frac{1}{|D|} \sum_{r \in D} Count_r(c) \qquad (1)$$

Where r labels a report in the collection, and $Count_r(c)$ is the number of times the concept c appears in report r.

This scoring function gives higher weight to concepts which many reports mentioned many times. These are, presumably, the terms considered important to the subject of interest. We ignore concepts (bigrams) composed entirely of stop words. To model the coverage of a report, we calculate a weighted sum of the concepts it mentions (multiple mentions do not increase this score), using the prevalence score as the weight, as shown in Equation 2.

$$CoverScore(r \in D) = \sum_{c \in Concepts(r)} prev_D(c) \qquad (2)$$

Here, $Concepts(r)$ is the set of concepts appearing at least once in report r. The system produces a ranking of the reports in order of their coverage score (where highest is considered best).

3.2 Evaluation

As a gold standard, we use the average of the scores given to each report by the human judges[3]. Since we are interested in ranking reports by coverage, we convert the scores from the original numerical scale to a ranked list. We evaluate the performance of the algorithms (and of the individual judges) using Kendall's Tau to measure concordance with the gold standard. Kendall's Tau coefficient (τ_k) is commonly used (e.g., Jijkoun and Hofmann 2009) to compare rankings, and looks at the number of pairs of ranked items that agree or disagree with the ordering in the gold standard. Let $T = \{(a_i, a_j) : a_i \prec_g a_j\}$ denote the set of pairs ordered in the gold standard (a_i precedes a_j). Let $R = \{(a_l, a_m) : a_l \prec_r a_m\}$ denote the set of pairs ordered by a ranking algorithm. $C = T \cap R$ is the set of concordant pairs, i.e., pairs ordered the same way in the gold standard and in the ranking, and $D = \overline{T} \cap R$ is the set of discordant pairs. Kendall's rank correlation coefficient τ_k is defined as follows:

$$\tau_k = \frac{|C| - |D|}{|T|} \qquad (3)$$

The value of τ_k ranges from -1 (reversed ranking) to 1 (perfect agreement), with 0 being equivalent to a random ranking (50% agreement). As a simple baseline system, we rank the reports according to their length in words, which asserts that a longer document has "more coverage". For comparison, we also examine agreement between individual human judges and the gold standard. In each scenario, we calculate the average agreement (Tau value) between an individual judge and the gold standard, and also look at the highest and lowest Tau value from among the individual judges.

3.3 Results

Figure 1 presents the results of our ranking experiments on each of the eight scenarios.

Human Performance There is a relatively wide range of performance among the human

of two reports on completely different subjects, based on external knowledge. For our usage scenario, this is not an issue.

[2]We also experimented with unigram and trigram representations, which did not do as well as the bigram representation (as suggested by Gillick and Favre 2009).

[3]Since the judges in the NIST experiment were also the writers of the documents, and the workshop report (Morse et al., 2004) identified a bias of the individual judges when evaluating their own reports, we did not include the score given by the report's author in this average. I.e, the gold standard score was the average of the scores given by the 6 judges who were not the author.

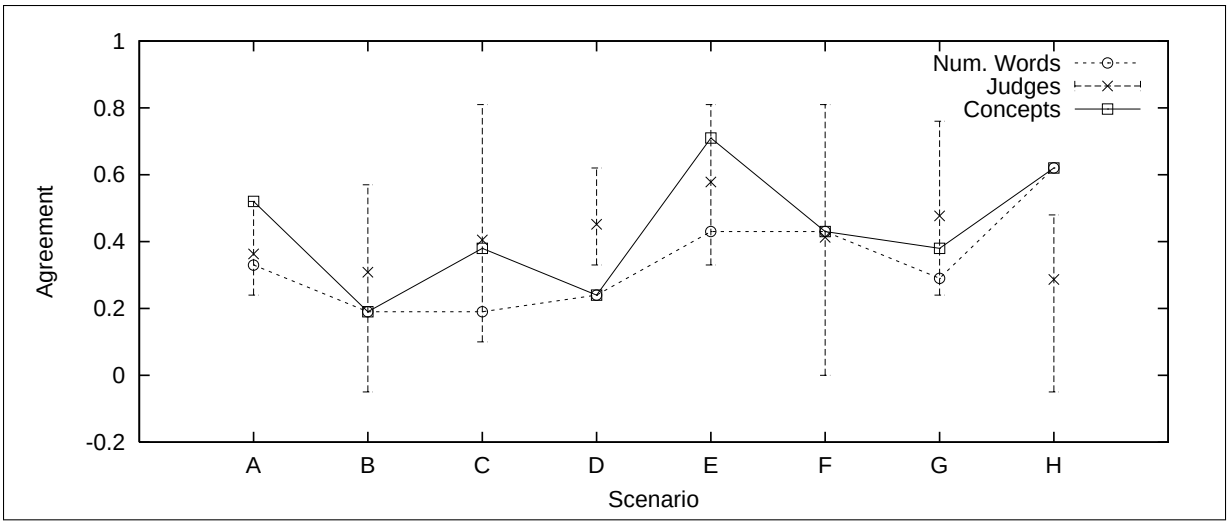

Figure 1: Agreement scores (Kendall's Tau) for the word-count baseline (Num. Words), the concept-based algorithm (Concepts). Scores for the individual human judges (Judges) are given as a range from lowest to highest individual agreement score, with 'x' indicating the average.

judges. This is indicative of the cognitive complexity of the notion of coverage. We can see that some human judges are better than others at assessing this quality (as represented by the gold standard). It is interesting to note that there was not a single individual judge who was worst or best across all cases. A system that outperforms some individual human judge on this task can be considered successful, and one that surpasses the average individual agreement even more so.

Baseline The experiments bear out the intuition that led to our choice of baseline. The number of words in a document is significantly correlated with its gold-standard coverage rank. This simple baseline is surprisingly effective, outperforming the worst human judge in seven out of eight scenarios, and doing better than the average individual in two of them.

System Performance Our concept-based ranking system exhibits very strong performance[4]. It is as good or better than the baseline in all scenarios. It outperforms the worst individual human judge in seven of the eight cases, and does better than the average individual agreement in four. This is in spite of the fact that the system had no access to the sources of information available to the writers (and judges) of the reports.

When calculating the overall agreement with the gold-standard over all the scenarios, our concept-based system came in second, outperforming all but one of the human judges. The word-count baseline was in the last place, close behind a human judge. A unigram-based system (which was our first attempt at modeling concepts) tied for third place with two human judges.

3.4 Discussion and Future Work

We have presented a system for assessing the relative quality of intelligence reports with regard to their coverage. Our method makes use of ideas from the summarization literature designed to capture the notion of content units and relevance. Our system is as accurate as individual human judges for this concept.

The bigram representation we employ is only a rough approximation of actual concepts or themes. We are in the process of obtaining more documents in the domain, which will allow the use of more complex models and more sophisticated representations. In particular, we are considering clusters of terms and probabilistic topic models such as LDA (Blei et al., 2003). However, the limitations of our domain, primar-

[4]Our conclusions are based on the observed differences in performance, although statistical significance is difficult to assess, due to the small sample size.

ily the small amount of relatively short documents, may restrict their applicability, and advocate instead the use of semantic knowledge and resources.

This work represents a first step in the complex task of assessing the quality of intelligence reports. In this paper we focused on coverage - perhaps the most important aspect in determining which single report to read among several. There are many other important factors in assessing quality, as described in Section 2.1. We will address these in future stages of the quality assessment project.

4 ACKNOWLEDGMENTS

The authors were funded by an IC Postdoc Grant (HM 1582-09-01-0022). The second author also acknowledges the support of the AQUAINT program, and the KDD program under NSF Grants SES 05-18543 and CCR 00-87022. We would like to thank Dr. Emile Morse of NIST for her generosity in providing the documents and set of judgments from the ARDA Challenge Workshop project, and Prof. Dragomir Radev for his assistance and advice. We would also like to thank the anonymous reviewers for their helpful comments.

References

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Burstein, Jill, Martin Chodorow, and Claudia Leacock. 2004. Automated essay evaluation: the criterion online writing service. *AI Mag.* 25:27–36.

Gillick, Dan and Benoit Favre. 2009. A scalable global model for summarization. In *Proc. of the Workshop on Integer Linear Programming for Natural Language Processing*. ACL, Stroudsburg, PA, USA, ILP '09, pages 10–18.

Gillick, Daniel, Benoit Favre, Dilek Hakkani-Tur, Berndt Bohnet, Yang Liu, and Shasha Xie. 2009. The ICSI/UTD Summarization System at TAC 2009. In *Proc. of the Text Analysis Conference workshop, Gaithersburg, MD (USA)*.

Goldstein, Jade, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz. 2000. Multi-document summarization by sentence extraction. In *Proc. of the 2000 NAACL-ANLP Workshop on Automatic summarization - Volume 4*. Association for Computational Linguistics, Stroudsburg, PA, USA, NAACL-ANLP-AutoSum '00, pages 40–48.

Haghighi, Aria and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proc. of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. ACL, Boulder, Colorado, pages 362–370.

Jijkoun, Valentin and Katja Hofmann. 2009. Generating a non-english subjectivity lexicon: Relations that matter. In *Proc. of the 12th Conference of the European Chapter of the ACL (EACL 2009)*. ACL, Athens, Greece, pages 398–405.

Larkey, Leah S. 1998. Automatic essay grading using text categorization techniques. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, New York, NY, USA, pages 90–95.

Morse, Emile L., Jean Scholtz, Paul Kantor, Diane Kelly, and Ying Sun. 2004. An investigation of evaluation metrics for analytic question answering. Available by request from the first author.

Nenkova, Ani, Lucy Vanderwende, and Kathleen McKeown. 2006. A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization. In *SIGIR*. ACM, pages 573–580.

Radev, Dragomir R., Hongyan Jing, Małgorzata Styś, and Daniel Tam. 2004. Centroid-based summarization of multiple documents. *Inf. Process. Manage.* 40:919–938.

Shermis, Mark D. and Jill C. Burstein, editors. 2002. *Automated Essay Scoring: A Cross-disciplinary Perspective*. Routledge, 1 edition.