ABSTRACT

Title of dissertation: STATISTICAL ANALYSIS OF AUTOMATIC SUMMARIZATION EVALUATION

Peter A. Rankel, Doctor of Philosophy, 2016

Dissertation directed by: Professor Eric V. Slud
Department of Mathematics
Dr. John M. Conroy
IDA/Center for Computing Sciences

This dissertation applies statistical methods to the evaluation of automatic summarization using data from the Text Analysis Conferences in 2008-2011. Several aspects of the evaluation framework itself are studied, including the statistical testing used to determine significant differences, the assessors, and the design of the experiment. In addition, a family of evaluation metrics is developed to predict the score an automatically generated summary would receive from a human judge and its results are demonstrated at the Text Analysis Conference. Finally, variations on the evaluation framework are studied and their relative merits considered. An over-arching theme of this dissertation is the application of standard statistical methods to data that does not conform to the usual testing assumptions.

STATISTICAL ANALYSIS OF AUTOMATIC SUMMARIZATION EVALUATION

by

Peter A. Rankel

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2016

Advisory Committee:
Professor Eric Slud, Chair/Advisor
Dr. John M. Conroy, Co-Chair/Co-Advisor
Professor Benjamin Kedem
Professor Paul J. Smith
Professor Doug Oard, Dean's Representative

Acknowledgments

There are many people who helped me along the way during my pursuit of this degree and I'd like to thank them for their kindness and support.

First and foremost, I'd like to thank my advisors, Eric Slud and John Conroy. Though they helped me in very different ways, they were both very patient and excellent researchers to work with. Dr. Conroy provided lots of suggestions, helped me learn Matlab, and made sure my results checked out. Dr. Slud asked many great questions about the methods I was using and made suggestions of how to do things better.

In addition to my advisors, I'd like to thank all of my other co-authors: Dianne O'Leary, Judith Schlesinger, Hoa Dang, Ani Nenkova, and Karolina Owczarzak. I enjoyed working with each of them and look forward to future opportunities for collaboration.

I'd like to thank the members of my committee for all the helpful suggestions and comments they made at the defense and in writing. I appreciated the feedback very much. I'd like to thank Dr. Kedem in particular for accepting my transfer from the pure MATH program to the STAT program so many years ago and arranging my remaining requirements in a very reasonable way. Both of these gestures were absolutely essential to my success in this program and I greatly appreciate them.

I'd like to thank everyone at Elder Research and at my previous employers. I greatly appreciate the flexibility they were able to give me so that I could continue working on my research and dissertation while working full-time.

My parents and family and friends have been supportive throughout this long endeavor and for that I am very grateful. Most of all, I appreciate the support they were able to provide to my family while I was away.

Finally, I'd like to thank my wife, Lauren, and our four children, Grace, Lillian, Katherine, and Henry for their unending patience and support. Thanks for being understanding of all my time away and for filling in during my absences.

Contents

List of Tables

List of Figures

List of Abbreviations

AESOP	Automatically Evaluating Summaries of Peers
ANOVA	Analysis of Variance
AQUAINT	(corpus of newswire text data)
ARDA	Advanced Research and Development Activity
ARE	Asymptotic Relative Efficiency
BEwT-E	Basic Elements with Transformations for Evaluation
BLEU	Bilingual Evaluation Understudy
CLASSY	Clustering, Linguistics, and Statistics for Summarization Yield
COLING	Conference on Computational Linguistics
DARPA	Defense Advanced Research Projects Agency
DUC	Document Understanding Conference
FASST-E	very Fast, very Accurate Sentence Splitter for Text – English
HB	Hybrid (from Monte Carlo simulation)
IDA	Institute for Defense Analyses
INEX	Initiative for the Evaluation of XML retrieval
LCS	Longest Common Subsequence
LSA	Latent Semantic Analysis
MAD	Median Absolute Deviation
MC	Monte Carlo (simulation)
NIST	National Institute of Standards and Technology
RHS	Right-hand side
ROSE	ROUGE Optimal Summarization Evaluation
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
SCU	Summary Content Unit
SEE	Summarization Evaluation Environment
TAC	Text Analysis Conference
VERT	Valuation using Enhanced Rationale Technique

Chapter 1

Introduction

The amount of information available in electronic form has seemed to be increasing without bound in the most recent decade. People today are facing a problem called Information Overload, and as a result, are looking for automated tools to help them process and understand vast amounts of documents. Examples of these tools include systems for information retrieval, machine translation, and the one focused on here, automatic summarization. The goal of automatic summarization is to condense a collection of (usually) related documents into a short form that includes the most important information. This information can then be more easily digested than the original (possibly large) documents. There are many such systems available and in use today, and they employ a wide array of advanced techniques to extract information and display it in a readable passage of text.

1.1 Motivation

Because multiple summarization systems do exist, users needing automatically generated summaries must decide which system is best suited for their particular needs. There are several ways to evaluate the effectiveness of a summarization system, and research into new methods is ongoing. The original method was ordinal scale human judgments, and is still the current gold standard. Automated methods have since been introduced, and some summarization workshops have added new tracks to further their research. These automated methods are often themselves evaluated by how well they agree with human scores.

1.2 The Problems

The recent summarization workshops at the National Institute of Standards and Technology have provided invaluable feedback for summarization researchers. The sets of news articles used at the workshop come from a range of story types and now comprise a comprehensive test set for evaluating algorithms that summarize news. In the current framework used for summarization evaluation, systems are compared using an unpaired t-test. Also, each summary is evaluated by one human only with no replication. This approach has various shortcomings that could be improved by enacting the changes suggested below.

This dissertation studies several aspects of this evaluation framework and suggests changes that could expand the impact of summarization evaluation. The data used in this research comes directly from the very summarization workshop whose framework I would like to improve.

In addition, the problem of how to create an automatic evaluation method is investigated by implementing a novel algorithm to predict the score a human would give to a summary. Details of this algorithm and its performance at the summarization workshop are included in Chapter 5.

1.3 Research Questions

The studies conducted here examine the existing framework of summarization evaluation. The framework itself is analyzed, challenges inherent in the problem are examined, and improvements are tested and suggested. The following research questions are addressed:

- In this framework, summarization systems are compared by their utility scores across multiple sets of documents. Is this testing arranged to allow for maximum statistical value? In particular, is the unpaired t-test the best choice for a test comparing two populations?

- Which features of a text summary can be used in different types of models to predict the score a human would use to rate the summary? Is it possible to create a family of automatic evaluation metrics, consisting of various types of statistical models, to predict the quality of a text summary, and if so, which features (co-variates) can be used in these models, and what will be their coefficients in the resulting model?

- Is it possible to measure the consistency of a human assessor, whose scores are the gold standard for summarization evaluation? What then is the effect on evaluation if the assessors are not completely consistent? How can the evaluation framework be made more robust to overcome these potential inconsistencies?

- What happens if certain changes to the way an evaluation metric is scored are introduced? In particular, what would be the effect on rankings if metrics were only judged on whether they agreed with human judgments on the question of significant differences between pairs of systems? What if evaluation metrics were only judged on how well they scored the top half of all summarization systems? Would a combination of several evaluation metrics perform better than any one individually? How would they be combined?

1.4 Statistical Overview

The data used in this experiment comes from summaries scored in several different ways. The scorings range from those assigned by humans to those automatically calculated using a machine. In each case, we have data that aligns with the document groups summarized by each system. Under this framework, it is possible to test different methods of comparing summarization systems. In particular, a comparison of paired and unpaired tests is conducted in Chapter 4 to see which has more power against alternatives of interest.

Using the text of summaries generated by machines or written by humans, several features can be calculated. These features, designed to capture linguistic or content information, are then entered into various different statistical models and combined to predict the score a human would assign to each summary. The predictions are then compared in Chapter 5 to the true human scores and Pearson correlation is used to measure accuracy.

Krippendorff's Alpha [30] is used to measure the consistency of human assessors. It is a simple measure of the average disagreement measured on each assessor individually. More details on this are included in Chapter 6.

Finally, in Chapter 7, the subject of statistical significance is investigated relative to the evaluation of evaluation metrics. In addition, metrics are combined in such a way that a difference was called significant only if all members of the combination reported significance individually. These combinations were measured according to several performance statistics.

1.5 Significance of this Study

Given the lack of rigorous statistical testing in the field of automatic summarization evaluation, there is a need to bring the potential improvement gained by using such methods. Misapplied statistical tests or tests that are not appropriate given the hypotheses required can lead to lost power and incorrect comparisons. This work motivates the need to use statistical testing in a more careful and thoughtful way in this area of research.

This dissertation directly yields the following contributions:

- A proposed adjustment to the evaluation procedure used at the Text Analysis

Conference and evidence that it would be more statistically sound and provide more robust comparisons (Chapter 4).

- Creation and validation of a family of automatic evaluation metrics specifically designed to correlate well with scores provided by humans (Chapter 5).
- An investigation into the effect of inconsistent assessors on the standard evaluation procedure, and suggestions for how to make the procedure less vulnerable to this effect (Chapter 6).
- Arguments and evidence supporting the claim that researchers should additionally report the significance level of the comparison between their system and the current state of the art. At the very least, this would help the community identify improvements that may be due to chance alone, and are unlikely to be reproducible (Chapter 7).

1.6 Chapter Overview

The remainder of this dissertation is organized as follows:

- Chapter 2 presents the history of the Summarization track at the Text Analysis Conference, where the data used in these experiments was generated. An overview of the evaluation framework is given, including concepts and terminology used in the summarization literature.
- Chapter 3 contains a discussion of the statistical assumptions placed on the data, including the intricate ways certain variables overlap. Unique characteristics of the data are explored, and expositions are given for the statistical tests and procedures used in later chapters.
- Chapter 4 reviews the way automatic summarization systems are ranked in a particular summarization task, and provides a suggestion for a more robust way

to compare systems. The suggested procedure is tested and shown to improve the evaluation framework.

- Chapter 5 introduces a novel family of algorithms for automatically predicting the quality of a text summary. The algorithms are based on features designed to measure content and linguistic quality, and combines these features using one of several regression methods.

- Chapter 6 investigates the consistency of human assessors, who play a critical role in summarization evaluation. The consistency of each assessor is measured, and its effect on the ranking of automatic summarization systems is discussed.

- Chapter 7 argues that the correct way to measure the effectiveness of an automatic content prediction algorithm is to determine how often it agrees with a gold standard assessment of whether two summarization systems have significantly different performance. Also presented are the results of an experiment to see if combinations of other evaluation metrics can do better on this enhanced task. Finally an explanation is given as to why it is important for researchers to report not just the ROUGE scores of their system, but also the degree to which their system is or is not significantly better than the previous state of the art.

- Chapter 8 sums up the contributions of these studies and considers potential avenues for future research in these areas.

Chapter 2

Background / The Data

In 2000, there were several government agencies (DARPA, ARDA, and NIST) interested in building powerful multi-purpose information systems. A group emerged from these agencies to focus on text documents, and in that year, the group hosted a two-day workshop called the Document Understanding Conference (DUC). The initial workshop was mostly for planning, but the DUC workshops in 2001-2007 included an evaluation series "to further progress in summarization and enable researchers [to] participate in large-scale experiments." In 2008, NIST expanded the scope of the workshop to include other research areas in text analysis. The new workshop was called the Text Analysis Conference (TAC) and included a track dedicated to summarization. The Document Understanding Conference essentially became a track at the Text Analysis Conference.

2.1 Text Analysis Conference: Summarization Task

The Summarization track was run at the Text Analysis Conference each year during 2008-2011, and again in 2014. The data used in this dissertation comes from the summarization tracks at the 2008-2011 TAC workshops. The specific tasks at the summarization tracks have evolved over the years. Variations have included the types and lengths of summaries created. For example, in the early years of DUC (2001-2004), the

Table 2.1: TAC Summarization track data during 2008-2011 (from [46])

Year	Document Sets	Automatic Summarizers	Human Summarizers
2008	48	58	8
2009	44	55	8
2010	46	43	8
2011	44	50	8

focus was generic summarization, and in DUC 2002, single-document summarization was done. Query-focused summarization was done in 2008-2009 and guided summarization was the focus during 2010-2011.

Some aspects of the summarization targets have remained consistent. For example, in most years, the number of document sets for multi-document summaries was approximately 45 (Table 2.1). Each document set normally contained twenty news articles about a particular topic, and the articles spanned a compact period of time. The twenty articles were then sorted by date and split into two groups of ten (first ten and last ten).

Starting at DUC 2007, and continuing through at least TAC 2011, participants were asked to create "update" summaries in addition to regular summaries. For the update portion of the summarization task, a short summary (perhaps 100 words) of the latter ten documents should be written under the assumption that the reader is already familiar with the content of the first ten documents.

2.2 Model Summaries

One way summaries are judged as effective information sources for users is to compare the content of the summaries to one or more human-written summaries of the same set of documents. In order for this to happen, there is a need for human-written summaries. In much of the DUC/TAC literature, these are referred to as "model summaries," since they serve as "models" for what a quality summary should look like. Detailed instructions are provided to the NIST associates who write these summaries in hopes that passages of uniform quality are produced. The instructions begin by describing the information need of the hypothetical user. Model summary writers are to assume such a user is an educated, adult, U.S. native who is aware of current events as they appear in the news. This user has been watching a news story develop over time and has subscribed to a news feed that sends articles from various sources as they are written. However, after the user falls behind and tries to get caught up, he becomes bothered that so many of the articles are repeating the same information. He would like to read summaries that highlight the content that is new and different.

The user initially will provide a topic statement expressing his information need. Relevant news articles arrive in batches over time, and for each batch, a summary will be written with the topic statement in mind. One hundred word summaries are to be written for both the first ten documents and the second ten. The initial summary is a standard topic-focused summary, but for the second summary, the writer should assume the user is aware of all the information contained in the first ten documents. The topic statement is still the same, but now the user is interested in only the new information that addresses the original information need. [68]

2.2.1 Scoring Instructions for Update Summarization

The announced procedure for scoring a summary was that it would first be truncated to 100 words. In order to evaluate a summary with the Pyramid Method (Section 2.3.3), human-written summaries (or “model summaries”) must be written first. Each topic statement and its two document sets were given to four different NIST volunteers. For each document set, the volunteer will create a 100 word summary that addresses the information need expressed in the topic statement [68]. Assessors were guided by the following instructions [67]. In addition to the Pyramid evaluation, the assessor will give an overall responsiveness score to each summary. This score will reflect both content and readability and will be judged according to the scale in Table 2.2:

Table 2.2: Meaning of scores assigned to summaries.

2008, 2010-2011	2009	Meaning
1	1-2	Very Poor
2	3-4	Poor
3	5-6	Barely Acceptable
4	7-8	Good
5	9-10	Very Good

2.3 Human Measures

After a contributed summary gets produced, an assessor working for NIST scores the summary according to several criteria. The first two scores, **overall responsiveness** and **overall readability** (or linguistic quality) are assigned by a judge as a single integer between 1 and 5 inclusive (except for 2009, where the scores were integers between 1 and 10) according to the values in Table 2.2. These first two scores are very coarse, considering how many factors a judge has to consider before ending up at a single integer. If a judge thinks a summary's score falls close to the middle of two consecutive integer scores, he or she has no choice but to choose the score believed to be closer. These Likert-style scores are intended to be strictly ordinal, but can certainly be interpreted as "ratio" or "interval." It is most likely not the case that a score of 2 can be thought of as being twice as good as a score of 1, since an empty summary or one with no coherent sentences or phrases still receives a 1, so the scores are not "ratio" in their raw form. However, if the scores are shifted from $1-5$ down to $0-4$, now it is possible that a 2 is twice as good as a 1, leading to a quasi-ratio interpretation. In fact, assessors have repeatedly requested that the scores be shifted in this way, although it not certain if the requests are inspired by this line of reasoning.

Much of the analysis done here is done assuming the scores follow an "interval" interpretation. Even though the judges are not instructed to give scores in this way, it is natural to think of the scores as having once been a continuous value that was rounded to the nearest possible score. The third and final score given to a summary is called the

Pyramid score (Section 2.3.3), and is calculated in a way that resembles a rubric.

2.3.1 Overall Responsiveness

A NIST assessor gives an overall responsiveness score to each summary. The overall responsiveness score is based on both content and readability/fluency. Its intention is to reflect the degree to which a summary is responding to the information need expressed in the topic statement, considering its informational content as well as linguistic quality.

2.3.2 Readability / Linguistic Quality

Each summary is also scored purely on its readability/fluency, without regard to whether it contains any information that responds to the topic statement. This score is based on factors such as the summary's grammaticality, non-redundancy, referential clarity, focus, and structure and coherence. This score requires no comparison with a gold standard summary and is unaffected by the topic being summarized. In years prior to 2008, each of the factors contributing to the readability score was evaluated on a separate 1-5 scale. This method of scoring was replaced by a single score in the range 1-5, but the individual factors are still considered. The details of how each factor would be graded separately [10] are included here:

1. **Grammaticality**: The summary should have no datelines, system-internal formatting, capitalization errors or obviously ungrammatical sentences (e.g., fragments, missing components) that make the text difficult to read.

2. **Non-redundancy**: There should be no unnecessary repetition in the summary. Unnecessary repetition might take the form of whole sentences that are repeated, or repeated facts, or the repeated use of a noun or noun phrase (e.g., "Bill Clinton") when a pronoun ("he") would suffice.

3. **Referential clarity**: It should be easy to identify to whom or what the pronouns and noun phrases in the summary refer. If a person or other entity is mentioned,

it should be clear what its role in the story is. So, a reference would be unclear if an entity is referenced but its identity or relation to the story remains unclear.

4. **Focus**: The summary should have a focus; sentences should only contain information that is related to the rest of the summary.

5. **Structure and Coherence**: The summary should be well-structured and well-organized. The summary should not just be a heap of related information, but should build from sentence to sentence to a coherent body of information about a topic.

2.3.3 Pyramid

The pyramid score [43] is designed to measure how closely a candidate summary's information content matches the content in one or more human-written summaries. The idea of evaluating summaries by counting content units matched with reference summaries dates back to DUC 2001, when software known as SEE (Summarization Evaluation Environment [12, 45]), assisted human evaluators in keeping track of exactly this. The pyramid score starts with (usually four) reference summaries and creates a distribution of content that would be expected in a summary. As one would expect, human summarizers do not produce identical summaries even when starting with the same set of source documents. They produce summaries that partially overlap. In theory, the most important information is captured in every human summary, but if the summary length cannot accommodate it all, a judgment call will have to be made as to what is cut. If there is extra room in the summary after including the most important information, then another judgment has to be made about which of the less important information is included. Each time a human makes a judgment, the likelihood that differing summaries will be produced increases. The pyramid method counts the frequency of occurrence of each information nugget in each human summary and uses those frequencies to score candidate summaries.

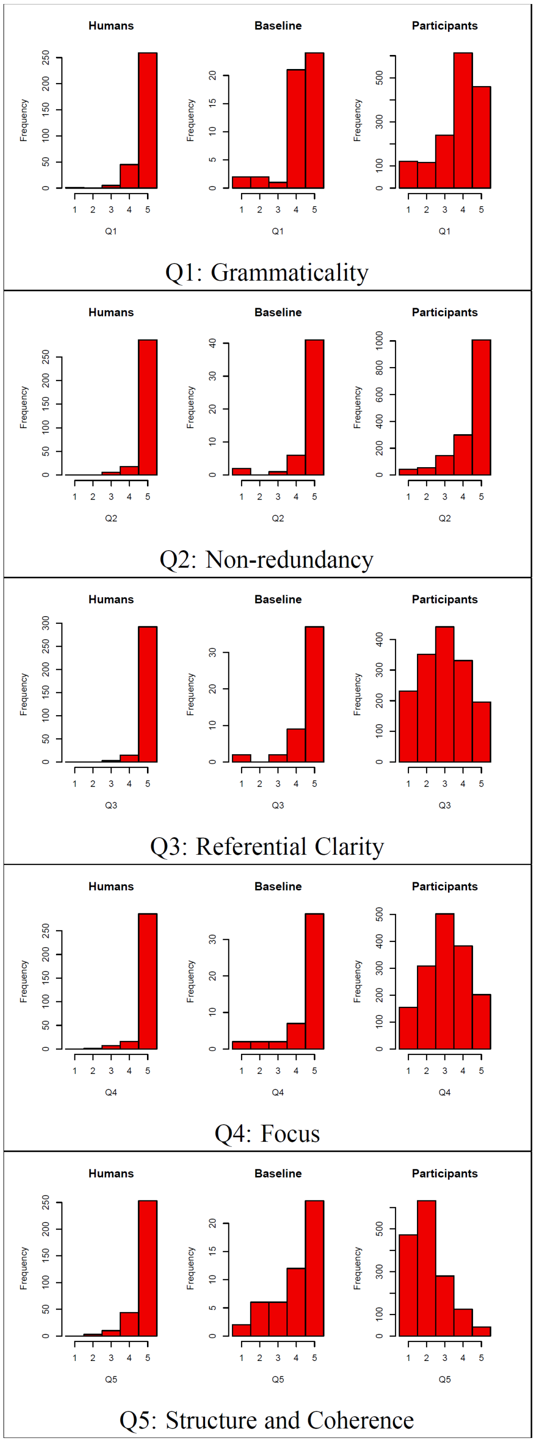


Figure 2.1: Distribution of scores for each of the five components of the linguistic/readability score (from [10]).

2.3.3.1 Summary Content Units (SCUs): Definition and Illustration

The individual pieces of information whose frequencies are counted by the pyramid method are called *summary content units* (or SCUs). These SCUs are annotated in the human summaries in order to match up information contained in multiple summaries. The annotator assigns a label to each SCU in each human summary and then counts the frequencies of each SCU across all human summaries. The following is an example (taken from [53]) of an SCU contained in every human summary, but each time expressed with a different phrase. The weight of this SCU is 4:

SCU 13 (W=4) Plaid Cymru is the Welsh nationalist party

C1 Plaid Cymru, the Welsh nationalist party

C2 the Welsh nationalist party, Plaid Cymru

C3 Plaid Cymru, the Welsh nationalist party

C4 Wales Nationalist Party (Plaid Cymru)

The annotator has some flexibility in choosing the label for the summary content unit, but should always aim for the label to be a paraphrase from each of the model summaries' contributing texts. The contributing texts can sometimes be less explicit than, or slightly different from, the label supplied by the annotator. The label is created to capture the shared meaning of each of the supporting texts' phrases. The following is an example illustrating that text can sometimes be segmented to form a contributing phrase:

SCU 49 (W=4) Plaid Cymru wants full independence

C1 Plaid Cymru wants full independence

C2 Plaid Cymru...whose policy is to...go for an independent Wales within the European community

C3 calls by...(Plaid Cymru)...fully self-governing Wales within the European Community

C4 Plaid Cymru...its campaign for equal rights to Welsh self-determination

If it is the case that quantities, or units of measure, are the only thing keeping phrases from being precisely equal, some liberties can be taken to craft the label in a way that includes as many phrases as possible. The label would then capture the annotator's opinion that the quantity in question was not required to be exactly equal to convey the same general idea. The following example illustrates this property:

SCU 77 (W=4) Wales has about 3 dozen district councils

C1 37 districts in Wales

C2 37 district councils

C3 38 Welsh districts

C4 37 district councils,

2.4 Evolution of the Summarization Track at DUC/TAC

The flavor of summarization tasks changed somewhat during the years in which our data was collected (2008-2011). The following sections contain notes detailing the evolution of summarization tasks during this time period.

2.4.1 Tighter Chronology and an Expanded Scale

In 2008, there was much variation in the amount of time elapsed between the first and last articles of a document set. In 2009, more of an effort was made to find collections of topic-focused documents that occurred closer together in time.

Another change for 2009 was that overall responsiveness, the main evaluation measure of automatic summarization, began to be graded on a 10-point scale instead of the

5-point scale used up to this point. Grading on a 5-point scale can be very difficult, especially considering that the lowest grade (1) corresponds to “very poor” summaries, while the highest grade (5), just a few points higher was given to “very good” summaries. While the 10-point scale did give reviewers more room to (discriminate/demarcate/delineate) the summaries of vastly different quality, it also created a small compatibility issue for algorithms designed to handle more than one year of data, since one of the scales would have to be translated into the other using any of a number of possible formulas. One way to turn a 10-point scale into a 5-point scale is to map the top score to the top score and the bottom score to the bottom score. The following formula sends 10 to 5 and 1 to 1: $y = \frac{5}{9}(x-1)+1$. Of course this has the unfortunate effect of turning the other eight scores from integers into non-integer fractions. The other possible translations have their pros and cons as well.

2.4.1.1 Automatically Evaluating Summaries Of Peers (AESOP)

Possibly the biggest change announced in 2009 was the creation of a new task related to summarization. The new task was called Automatically Evaluating Summaries Of Peers (AESOP) and the goal was to ascertain the quality of a text summary in an automated way. More specifically, the goal was to predict the score a summary would receive when it was scored by a human in either overall responsiveness or in pyramid score. The driving force behind the creation of this task was the potential to build systems that could automatically evaluate summaries. If built, these could then support the development of summarization systems.

2.4.2 Guided Summarization

A new flavor of summarization became the focus of the Summarization track at TAC in 2010. It was called “guided summarization” and its goal was to “encourage a deeper linguistic (semantic) analysis of the source documents instead of relying only on document word frequencies to select important concepts.” The basic task of writing a 100-word summary of ten newswire documents from a single topic stayed the same in 2010, but this time participants were asked to have their summaries address a list of pre-defined “aspects.” In addition, the topics governing the sets of documents themselves came from a broad array of “categories.” The list of aspects per category would be the same across different document sets within a category. Examples of each were as follows [68]:

1. **Accidents and Natural Disasters**: what happened; date; location; reasons for accident/disaster; casualties; damages; rescue efforts/countermeasures

2. **Attacks**: what happened; date; location; casualties; damages; perpetrators; rescue efforts/countermeasures

3. **Health and Safety**: what is the issue; who is affected; how they are affected; why it happens; countermeasures

4. **Endangered Resources**: description of resource; importance of resource; threats to resource; countermeasures

5. **Trials and Investigations**: who is under investigation, who is investigating/suing; why (general); specific charges; sentence/consequences; how do they plead/react to charges

Another part of the summarization track that stayed the same in 2010 was the division of each topic’s twenty documents into two sets of ten. The first ten documents were summarized according to the aspects mentioned above, and the final ten contributed to a “guided update summary.” Both summaries were 100 words and were evaluated for readability, content, and overall responsiveness. Just as in 2009, the AESOP task was also run in 2010.

2.4.3 More Focused Model Summaries

In 2011, guided summarization remained the central theme of the summarization track. However, a new explanation was given for its continued focus. The problem guided summarization was intended to solve in 2011 was the absence of a single "gold standard" that automatic systems can model. Prior to providing this guidance, model summaries created by humans could still vary quite a bit, even when created using the same ten source documents. This created a notion of "importance" of certain facts and could be approximated by counting the number of model summaries containing each fact.

In order to end up with more uniform model summaries that each contained a higher percentage of identical facts, topics were chosen from template-like categories that contained highly predictable elements. In addition, the human summarizers were instructed to create summaries following the same guidelines as the machines. The goal of all this was a unified information model that automatic summarizers could emulate.

The 2011 track kept most other things the same, including the 100-word length and initial and update summaries of ten documents each. However, further information was given on the interplay between "guided" and "update" summarization. It was explicitly stated in the 2011 instructions that the "update" portion of the summarization took precedence over the "guided" portion. Hence, it was more important to not repeat information, even if it meant not including information for each aspect of the category. The 2011 data was the first to not come from AQUAINT or AQUAINT-2. It instead came from the newswire portion of another track at TAC 2010. The collection spanned the years 2007-2008 and consisted of documents taken from the New York Times, the Associated Press, and the Xinhua News Agency (English language) newswires.

Each summary should cover all the aspects relevant to its category, and it may contain other relevant information as well. The categories, their aspects, and their numerical IDs are as follows [68]:

1. Accidents and Natural Disasters:

   1.1 WHAT: what happened

   1.2 WHEN: date, time, other temporal placement markers

   1.3 WHERE: physical location

1.4 WHY: reasons for accident/disaster

1.5 WHO_AFFECTED: casualties (death, injury), or individuals otherwise negatively affected by the accident/disaster

1.6 DAMAGES: damages caused by the accident/disaster

1.7 COUNTERMEASURES: countermeasures, rescue efforts, prevention efforts, other reactions to the accident/disaster

2. Attacks (Criminal/Terrorist):

   2.1 WHAT: what happened

   2.2 WHEN: date, time, other temporal placement markers

   2.3 WHERE: physical location

   2.4 PERPETRATORS: individuals or groups responsible for the attack

   2.5 WHY: reasons for the attack

   2.6 WHO_AFFECTED: casualties (death, injury), or individuals otherwise negatively affected by the attack

   2.7 DAMAGES: damages caused by the attack

   2.8 COUNTERMEASURES: countermeasures, rescue efforts, prevention efforts, other reactions to the attack (e.g. police investigations)

3. Health and Safety:

   3.1 WHAT: what is the issue

   3.2 WHO_AFFECTED: who is affected by the health/safety issue

   3.3 HOW: how they are affected

   3.4 WHY: why the health/safety issue occurs

   3.5 COUNTERMEASURES: countermeasures, prevention efforts

4. Endangered Resources:

4.1 WHAT: description of resource

4.2 IMPORTANCE: importance of resource

4.3 THREATS: threats to the resource

4.4 COUNTERMEASURES: countermeasures, prevention efforts

5. Investigations and Trials (Criminal/Legal/Other):

5.1 WHO: who is a defendant or under investigation

5.2 WHO_INV: who is investigating, prosecuting, or judging

5.3 WHY: general reasons for the investigation/trial

5.4 CHARGES: specific charges to the defendant

5.5 PLEAD: defendant's reaction to charges, including admission of guilt, denial of charges, or explanations

5.6 SENTENCE: sentence or other consequences to defendant

The categories and aspects for 2011 were developed based on model summaries from previous DUC and TAC summarization tasks. Rather than prescribing what information should or should not be included in a summary about a specific category, previously written model summaries were studied to see which aspects were most often captured. Examples of these model summaries which have been annotated with the above aspects are included in Appendix B.

A small change was made to the AESOP task in 2011. In addition to metrics that reflect summary content (Pyramid, Responsiveness), it was decided that AESOP would target Readability in 2011. Also, in addition to reporting which AESOP systems were able to best correlate with system performance, correlations at the summary level (within each topic) were also reported.

The final change in 2011 worth mentioning here was the introduction of a multi-lingual summarization task. The data from this particular task was not used in any of the experiments mentioned here, but for completeness, a brief description of the task follows. According to its website, the MultiLingual task aimed to “evaluate the application of (partially or fully) language-independent summarization algorithms on a variety of languages.” Participants were asked to develop systems capable of producing summaries in a range of languages. The specific task was to generate a fluent, representative summary from a set of documents in any of Arabic, Czech, English, French, Hebrew, Hindi, Greek. All documents in a set would be from a single language, and the generated summary should be from the same language. Summaries for this task should be between 240 and 250 words.

2.5 Automatic Measures

Participants in the AESOP task at TAC 2009-2011 attempted to create an algorithm that could automatically score a text summary. These systems were rated on how closely the scores they generated compared to the scores given by human annotators. One of the earliest systems to achieve success in this area was called ROUGE. Other examples of automatic measures are described in Chapter 5.

2.5.1 ROUGE

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [34] is a package for the automatic evaluation of summaries. It is comprised of ROUGE-n (for n a natural number), ROUGE-L, ROUGE-W, ROUGE-S, and ROUGE-BE (basic elements). The basic idea of these ROUGE variants (except for ROUGE-BE) is to count the number of n-grams in a candidate summary that match an n-gram in the reference summaries. It is reminiscent of the Pyramid method but is fully automatic and uses matching n-grams to serve as a proxy for Pyramid’s use of matching content units. ROUGE grew out of BLEU (BiLingual Evaluation Understudy) [48], an automatic, n-gram based method for evaluating machine translations.

2.5.1.1 ROUGE-n

The formula for ROUGE-n is as follows:

$$\text{ROUGE-}n \;= \frac{\sum_{S\in\{ReferenceSummaries\}}\sum_{gram_n\in S} Count_{match}(gram_n)}{\sum_{S\in\{ReferenceSummaries\}}\sum_{gram_n\in S} Count(gram_n)},$$

where the evaluation is done on sequences of n consecutive words. The comparison is done with each reference summary individually, and the maximum match percentage is reported as the score. Although ROUGE-n can be calculated for any positive number n, the variants of ROUGE-n used most often are ROUGE-1 and ROUGE-2.

2.5.1.2 ROUGE-L

The other variants of ROUGE are not used as frequently, but occasionally provide scores that correlate well with human-provided scores. ROUGE-L is a measure of the longest common subsequence between sentences in a candidate summary and those in a reference summary. The score is a composite of the ratio of the length of the longest common subsequence to both the length of the candidate sentence and length of the reference sentence. There are two advantages of ROUGE-L. One is that LCS does not require consecutive matches. It only measures in-sequence matches that can reflect sentence level word order as n-grams. The other advantage of ROUGE-L is that it takes into account n-grams of different lengths without the need to specify n in advance.

2.5.1.3 ROUGE-W and ROUGE-S[n]

ROUGE-W is a variant of ROUGE-L that adds weighting to the longest common subsequence. ROUGE-S measures “skip-bigram co-occurrences.” Hence it is similar to ROUGE-2 because it focuses exclusively on bi-grams, but this version allows the bigrams to be non-consecutive. An extension of ROUGE-S called ROUGE-SU also counts unigram matches between candidate sentences (with the “U” in ROUGE-SU commonly read as “up to”). A popular candidate in this family is ROUGE-SU4, which counts bigrams co-occurrences separated by a maximum of four intermediary words, and also counts unigram co-occurrences.

Chapter 3

Statistical Methods

In this chapter, I discuss many of the statistical methods used in later chapters for the analysis of summarization evaluation. In some cases, I have included an extra discussion about how the methods can be adapted to summarization data. The first few subsections describe statistical tests for testing whether or not two populations are different, but the very first section describes the general format of the data and assumptions therein.

3.1 General Assumptions of the Data

The data used in this dissertation can be nicely represented by the color plot in Figure 3.1. The colors in that plot refer to the scores received by summaries according to some pre-specified criteria (ex. pyramid score, overall responsiveness, etc.). The scores are organized into a matrix such that each row represents a particular topic and each column represents the person or machine that summarized the documents from the given topic. In addition to the topic, each row contains other information that affects its scores. Each topic comes from one of five different categories, and the summaries from each topic are scored by a particular person. Each of the five categories is represented in the list of topics, and each of eight different assessors is responsible for producing entire rows of individual scores.

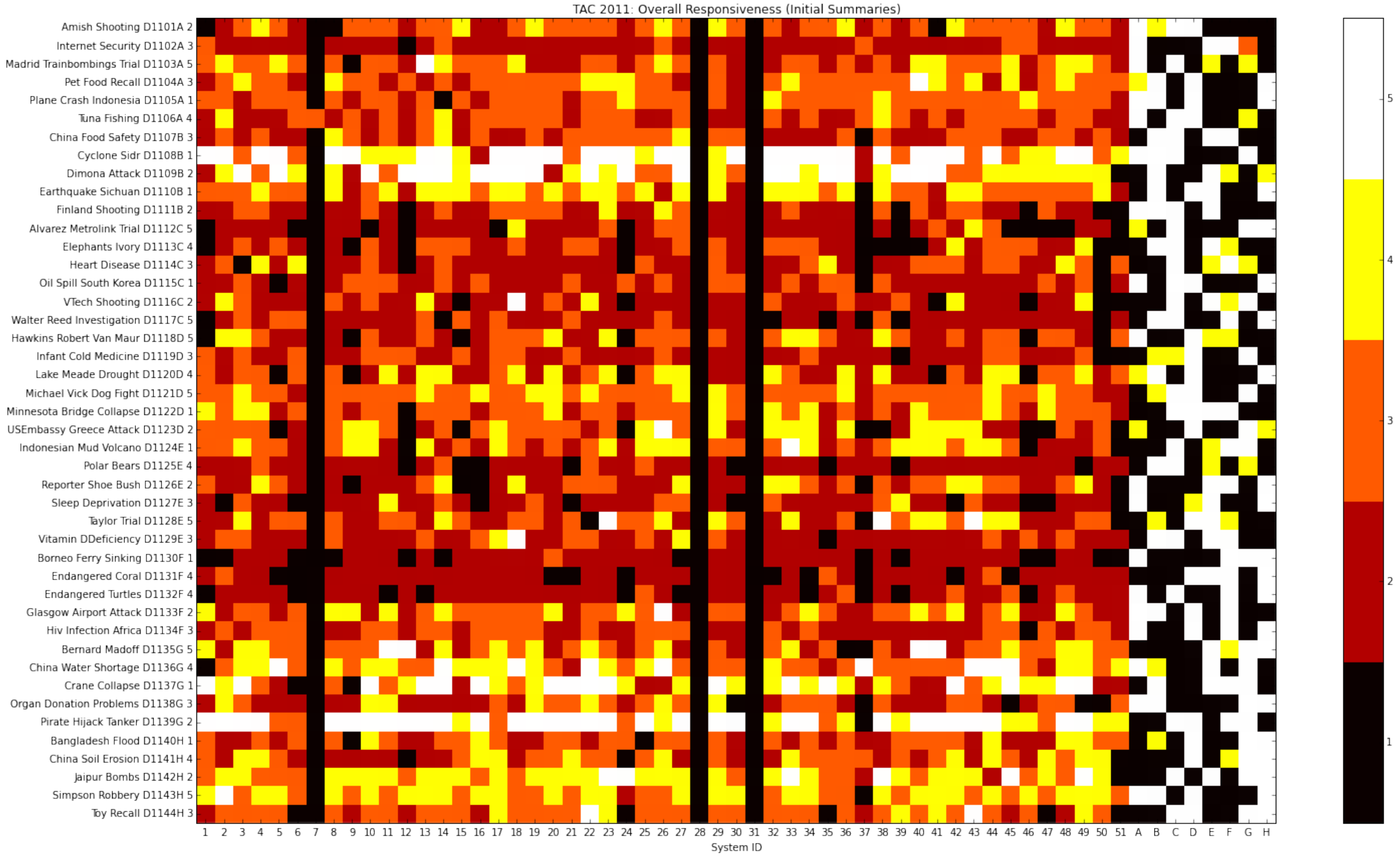


Figure 3.1: Overall responsiveness scores for the TAC 2011 non-update summary task, organized by document set (y-axis) and summarizer (x-axis). The 59 summarizers fall into two distinct groups: machine systems (columns 1–51) and humans (columns A–H). Note that each human only summarized half of the document sets, thus creating 22 missing values in each of columns A–H. Black is used to indicate missing values in columns A–H and low scores in columns 1–51. Rows are labeled with topic name, document ID (in which the last character denotes the assessor), and category (numbered 1–5: Table 3.1 lists categories in numerical order).

In the simplest possible scenario, one would assume that the rows and columns of this matrix are identically distributed. If this were true, we could model each score Y in the matrix as coming from a single distribution. Our main interest here is comparing the systems that summarize the documents. We are testing for differences between systems, and hence our null hypothesis is that the systems (and hence the columns) are not different. In this way, all of the variation we include in the model will be due to differences among the rows. If the columns were in fact identical, then the only variation in scores would be due to the ways in which the rows differed from one another.

The rows of the matrix differ in several ways. Each row represents a specific topic from which documents were collected and summarized, and each topic is different. However, each topic comes from one of the five categories listed in Table 3.1, so there is a natural grouping of topics. In addition, each row of summaries is judged by one of eight different assessors, labeled as A-H in Table 3.1. I would have liked to include a topic effect in the model, but the variables "assessor" and "topic" are not separately identifiable. For any given topic, every summary was scored by the same assessor, so we cannot tell, given our data, whether the scores in a row were higher due to the assessor or due to the topic being summarized.

The situation is not quite as bad between assessor and category. Here, Table 3.1 gives the counts of how many times each assessor scored summaries from topics from each of the five different categories. The two variables are almost completely crossed, in that almost every assessor worked on every category. Only three combinations of assessor and category are missing from our data. In a few cases, we see a combination occurring more than once.

Within the above framework, we arrive at the following model:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk},$$

where i indexes assessor (A-H), j indexes category (1-5), and k indexes replications within combinations of assessor and category. In the model, we assume that the ϵ_{ijk}

Table 3.1: Number of times each assessor scored summaries from each of the five different categories.

	Assessor							
Category	A	B	C	D	E	F	G	H
Accidents and Natural Disasters	1	2	1	1	1	1	1	1
Attacks	1	2	1	1	1	1	1	1
Health and Safety	2	1	1	1	2	1	1	1
Endangered Resources	1	0	1	1	1	2	1	1
Trials and Investigations	1	0	2	2	1	0	1	1

variables are independent and identically distributed.

The interaction effects γ_{ij} between assessor i and category j are assumed to be 0, since we have very little information about them. In particular, Table 3.1 specifies the small number of cases where we have any replication at all between an assessor and category combination. The α_i and β_j effects are considered fixed effects, and are subject to the side conditions $\sum_i \alpha_i = \sum_j \beta_j = 0$. Our interest here is whether all the α_i are equal in the presence of unknown β_j. The equation we end up with is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}.$$

In the sections below, we will use the notation $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ to represent two vectors of scores from the above matrix. The pair (X_i, Y_i) denotes the scores earned by summarization systems X and Y upon attempting to summarize document set i.

3.2 Wilcoxon Signed-Rank Test

The Wilcoxon signed-rank test [75] is designed to test whether the median of a symmetric population is a specific value θ_0. In the case of comparing two summarization systems, X and Y, the ordered pairs (X_i, Y_i) are *i.i.d.* but for each value of i, X_i and Y_i are not necessarily independent because X_i and Y_i are scores for summaries of the same document set. As we will see later, it is not the case that all document sets are equally difficult to summarize.

Given the X and Y data, the first thing we do in this paired context is compute $Z_i = Y_i - X_i$. Then, in order to test whether $P(X > Y) = P(X < Y)$, we use the signed-rank test and test whether the median of $Z_1, \ldots, Z_n$ is 0. One additional assumption for this test is that the input data are continuous, such that $P(Z_i = 0) = 0$. We will often violate this last assumption, and will consider the effects of that violation later.

The null hypothesis we test here is

$$H_0 : (X, Y) \overset{d}{=} (Y, X),$$

where the symbol $\overset{d}{=}$ means "equal in distribution." The random variables X and Y are called "exchangeable" when H_0 is true.

The alternatives against which we seek power can be described differently depending on our assumptions about the document sets being summarized. If we make the simplifying assumption that the document sets are identically distributed, i.e., equally easy or difficult to summarize, then the X and Y vectors we're comparing become completely independent. In this situation, we seek power against the stochastically ordered alternatives. These are any alternative that takes the form:

$$H_A : P(Y \geq t) \leq P(X \geq t) \tag{3.2.0.1}$$

for all t, with strict inequality for some t.

Due to the intricacies of human language, it is much more realistic that the document sets are not identically distributed. Now the situation becomes more complex because the X and Y vectors are no longer independent. Here the alternatives against which we seek power take the form

$$P(Z_1 + Z_2 > 0) > P(Z_1 + Z_2 < 0) \tag{3.2.0.2}$$

A consequence of the null hypothesis mentioned above is that each Z_i comes from a distribution symmetric about 0. The Wilcoxon signed rank test statistic is then defined as:

$$W = \sum_{i=1}^{n} \psi_i R_i$$

where R_i is the rank of $|Z_i|$ among $|Z_1|, \ldots, |Z_n|$, and

$$\psi_i = \begin{cases} 1, & \text{if } Z_i > 0, \\ 0, & \text{if } Z_i < 0. \end{cases}$$

Since $E(W) = \frac{n(n+1)}{4}$, we reject H_0 when W is sufficiently far from this value, as determined by tabulated values available in, for example, [23]. Note that W can never be less than 0 or more than $\frac{n(n+1)}{2}$. Under H_0, the variance of W is $\frac{n(n+1)(2n+1)}{24}$, (for a derivation of this fact, see [33]). As n goes to ∞, W becomes asymptotically normally distributed (see Appendix A).

3.2.1 Ties

There are two types of ties that can occur when using the Wilcoxon signed-rank test. The first type occurs when $X_i = Y_i$ for some i. This type causes Z_i to be 0, and hence will be referred to as a "zero tie" (often just called "zeros" in the literature, especially when viewing the test as a one-sample test on the differences). There are several ways of handling this type of tie, and the simplest is to discard the zero values and reduce the value of n accordingly. This was Wilcoxon's own advice [76]. Another method for handling zeros was suggested by Pratt [51], and involved dropping the ranks that belong to the zeros. Pratt's procedure is used in the test's description in books such as those by Hájek and Šidák [22] and Noether [44]. A third method of handling ties at zero is discussed by Putter in [52], and involves randomly assigning signs to the ranks belonging to the zeros. Still other methods are mentioned in Hájek and Šidák in [22]. All of these methods are reviewed in Conover [4]. For our purposes here, we follow Wilcoxon's advice and drop the values where $Z_i = 0$. In essence, we are conditioning on the event of no ties, which allows us to derive a reference distribution where the pairs are *i.i.d.* and conditioned on the reduced sample size.

The second type of tie is called a "non-zero tie" and occurs when $X_i - Y_i = X_j - Y_j \neq 0$ for some i and j, and consequently, $Z_i = Z_j \neq 0$. In this scenario, a typical procedure is to assign each of the observations in the tied group the average of the integer ranks that would have been assigned to the individual values in the tied group. For example, if the values listed in increasing order of absolute value are

$$1, 2, \mathbf{4}, \mathbf{4}, \mathbf{\text{-}4}, \mathbf{4}, 6, 7, -10, 13, -14,$$

then the corresponding ranks (before accounting for ties) would be

$$1, 2, \mathbf{3}, \mathbf{4}, \mathbf{5}, \mathbf{6}, 7, 8, 9, 10, 11.$$

After accounting for ties, the ranks assigned to all Z's with magnitude 4 would be the average of $3, 4, 5$, and 6. The final ranks would be

$$1, 2, \mathbf{4.5}, \mathbf{4.5}, \mathbf{4.5}, \mathbf{4.5}, 7, 8, 9, 10, 11.$$

When there are nonzero ties in the data, the large-sample testing procedure needs to be adjusted. Although the expected value of W stays the same, the variance under the null hypothesis becomes

$$\mathrm{var}(W) = \frac{1}{24}\left[n(n+1)(2n+1) - \frac{1}{2}\sum_{j=1}^{g} t_j(t_j-1)(t_j+1)\right],$$

where g denotes the number of tied groups of nonzero $|Z|$'s and t_j is the size of the jth tied group. In this sense, we take an untied observation to be a tied "group" of size 1. In the case where ties are possible, W once again has a limiting normal distribution (see Appendix A).

3.2.2 Consistency

The consistency of the Wilcoxon signed-rank test is discussed briefly in [23]. There, it is noted that the Wilcoxon signed-ranked test can detect a more general class of alternatives than the location-shift alternatives mentioned in most textbooks. Under the simplifying assumption mentioned in Section 3.2, in which documents are assumed to be identically distributed, this test is consistent against any alternative for which the median of Z_i is nonzero. This is because W suitably centered and scaled estimates $P(Z_i > 0)$ consistently. Hence, if $P(X > Y) \neq P(X < Y)$, then this test should reject with probability going to 1 as the sample size gets larger.

If an alternative like those found in (3.2.0.1) were true, then we would have $P(Z_i + Z_j > 0) > P(Z_i + Z_j < 0)$ (where Z_i, Z_j are *i.i.d.*). Because of the conclusions drawn in Section A.1.1 in Appendix A, we can conclude that this test is consistent against stochastically ordered alternatives of this form.

3.3 Paired and Unpaired t-test

The paired and unpaired t-tests are normally applied to data assumed to come from a continuous distribution that is normally distributed. For examples of how these get used in summarization data, see Section 3.7.1

3.4 Kruskal-Wallis Test[1]

The Kruskal-Wallis test is a generalization of the two-sample location test (in particular, it generalizes the Mann-Whitney-Wilcoxon Rank-Sum test) to situations in which the data consist of $k \geq 3$ random samples, one sample from each of k populations.

The data can be represented as a doubly-indexed array, X_{ij}, where, with our data, i ranges over the document collections and j varies according to the summarization systems. For each i, the $X_{i.}$ vector represents scores for a particular document set's summaries, and for each j, $X_{.j}$ represents all the scores for one system's summaries. This one-way layout is not necessarily balanced, so for each j, the corresponding sample has size n_j. In order to compute the Kruskal-Wallis test statistic, H, we combine all observations from the k samples and order them from least to greatest. Let r_{ij} denote

[1] Much of this description of the Kruskal-Wallis test was taken from [23]

the ranking of X_{ij} in the joint ranking and set

$$R_j = \sum_{i=1}^{n_j} r_{ij} \quad \text{and} \quad \overline{R_j} = \frac{R_j}{n_j}, j = 1, \ldots, k.$$

The null hypothesis is stochastic homogeneity of the columns, i.e., that $P(X_{ij} > X_{ik}) = P(X_{ij} < X_{ik})$ [72]. An equivalent form of the null hypothesis is equality of rank mean expected values, i.e., that the average rank assigned to each group is the same:

$$H_0 : E(\overline{R_1}) = E(\overline{R_2}) = \cdots = E(\overline{R_n}).$$

Under either of these equivalent null hypotheses, the k samples can be combined to form one larger sample from a single population. The alternative hypothesis is the general alternative of at least one population not stochastically equal to the union of the rest of the data. In general, the H test is consistent against any alternative hypothesis that implies this general property.

The Kruskal-Wallis test statistic H is then given by:

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} n_j \left(\overline{R_j} - \frac{N+1}{2} \right)^2,$$

where $N = \sum_{j=1}^{k} n_j$ and $(N+1)/2$ is the average rank assigned in the joint ranking.

To perform the test at the α level of significance, reject H_0 if $H \geq h_\alpha$, where the constant h_α is chosen to make the type I error probability equal to α. Values of h_α are tabulated for small values of $n_1, \ldots, n_k$.

3.4.1 Large-Sample Approximation

Under H_0, and the additional assumption of data coming from a continuous distribution, the test statistic H has, as $\min(n_1, \ldots, n_k)$ goes to infinity, an asymptotic χ^2 distribution with $k - 1$ degrees of freedom. The chi-square approximation of the above testing procedure is to reject H_0 when $H \geq \chi^2_{k-1,\alpha}$ [23].

3.4.2 Our Data

At first glance it would not appear that our data satisfy any of the assumptions listed above. For example, in the case of responsiveness and readability, the scores given to a summary are discrete integers between 1 and 5. As will be argued elsewhere, the document sets are not equally difficult to summarize, which would make the columns of X not act as independent samples from related distributions. And finally, it would be quite difficult for discrete distributions like these to be location shifts of one another.

However, there are ways in which our data can be thought of as approximating the above assumptions. For example, even though the rows of X are very different due to some document collections being more difficult to summarize, we can think of these collections as really being a random sample from the entire universe of possible document collections. In this sense, the elements of a column can be thought of as being independent and identically distributed according to some distribution dependent on the summarizer. The $1-5$ scores are not continuous in the basic sense, but they can be modeled as though they started out as a continuous variable and then were rounded to the nearest integer between 1 and 5. However, researchers have argued that with sufficient data, "discrete measures are often well-approximated by continuous distributions" [25]. In addition, even if the errors are not normal, tests such as the t-test are "relatively robust to many violations of normality" [25] (also [61]).

Absent these theoretical notions, one can always use something like the bootstrap to derive fully data dependent cutoffs. This is very often the simplest way to handle composite null hypotheses, as will be discussed further in Section 4.2. The basic problem is that with a composite null hypothesis, the pre-specified significance level may only be an upper bound for the attained size of the test. Using a cutoff derived from theoretical distributions, by large sample theory, one may end up with a different size test than anticipated. Resampling the data in a way that reflects its structure is a way to avoid this problem.

3.4.3 Ties

In the case of the Kruskal-Wallis test, these discrete values are unfortunately going to lead to a large number of tied ranks. As a particular example, take the responsiveness scores from the non-update portion of the summarization task at TAC 2011. The data start out as five discrete values 1, 2, 3, 4, 5, and end up as five different discrete values: 150.5, 688.5, 1427.5, 1961.5, 2282.5 (as seen in Table 3.2).

Table 3.2: Tied ranks derived from observed responsiveness scores from TAC 2011 non-update task.

Value	Tied Rank	Count	Percent
1	150.5	300	12.40%
2	688.5	776	32.07%
3	1427.5	702	29.01%
4	1961.5	366	15.12%
5	2282.5	276	11.40%

However, as in the case of the Wilcoxon signed-rank test, there is a procedure for dealing with ties in the Kruskal-Wallis test. In fact, the same method of assigning average ranks to each entry in a tied group is used, and the test statistic also gets altered in a similar way. Here, the new Kruskal-Wallis test statistic becomes

$$H' = \frac{H}{1 - \left(\sum_{j=1}^{g}(t_j^3 - t_j)/(N^3 - N)\right)},$$

where H is calculated using the rule for average ranks, g is the number of tied X groups, and t_j is the size of the jth tied group.

3.4.4 Consistency

Similar to the more general alternatives described for the signed-rank test, there is a more general *null* hypothesis for the Kruskal-Wallis testing procedure [23]. We can replace our original assumptions with the null hypothesis that all $N!/(\prod_{j=1}^{k} n_j!)$ assignments of n_k ranks to the treatment k observations are equally likely.

3.5 Regression Methods

For the three regression methods discussed in this section, the input data is an $n \times 13$ matrix A, consisting of columns of predictors, and an $n \times 3$ matrix B, consisting of columns b_1, b_2 and b_3 of the three human scores given to the summaries (overall responsiveness, pyramid score, and linguistic quality score). If we used standard linear regression (ordinary least squares), we would be looking to find $\hat{x}_i$ such that $||A\hat{x}_i - b_i||$ is minimized (where b_i is the particular column of human metrics we are trying to predict). Each of the three methods uses a variation on this theme.

3.5.1 Robust Regression

This method is finding $\hat{x}_i$ such that $||w(A\hat{x}_i - b_i)||$ is minimized, where w is a function that increases the weight of certain observations and decreases the weight of others. Here the function w is applied component-wise to the vector $A\hat{x}_i - b_i$. We used Matlab's `robustfit`, which follows an iterative re-weighting procedure. Its default weighting function, "bisquare", and default tuning constant, 4.685, were also used. "Bisquare" is defined as $w = (|r| < 1)(1 - r^2)^2$, where each operation is done component-wise.

The value r in the weight functions is $r = \text{resid}/(\text{tune} \times s\sqrt{1-h})$, where *resid* is the vector of residuals from the previous iteration, *tune* is a tuning constant that is divided into the residual vector before computing weights, h is the vector of leverage values from a least-squares fit, and s is an estimate of the standard deviation of the error term given by $s = \text{MAD}/0.6745$. Here MAD is the median absolute deviation of the residuals from their median. The constant 0.6745 makes the estimate unbiased for the normal distribution. If there are p columns in A, the smallest p absolute deviations are excluded when computing the median.

3.5.2 Non-negative Least Squares

This is essentially ordinary least squares, since we are trying to minimize $||A\hat{x}_i - b_i||$, but the variation here is that $\hat{x}_i$ is restricted to having non-negative entries. In order to avoid having features ignored, we first multiply all feature columns by ±1 to ensure each is positively correlated with the response vector. We used Matlab's `lsqnonneg`, which also uses an iterative procedure. The procedure starts with a set of possible basis vectors and computes the associated dual vector λ. It then selects the basis vector corresponding to the maximum value in λ and swaps out that vector in exchange for another possible candidate. This continues until each component of λ is less than or equal to zero, at which point a solution is reached.

3.5.3 Canonical Correlation

This method seeks a linear combination of the columns of A that has maximum correlation with a linear combination of the columns of B. As in [63], we can form the covariance matrix Σ of the matrix (A, B) and partition it as follows:

$$\Sigma = \begin{pmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{pmatrix}$$

where $\Sigma_{AB} = \Sigma'_{BA}$. Then, ρ^2, the first squared canonical correlation between A and B, is defined to be the maximum squared correlation between arbitrary linear combinations of the columns of A and B, say $v'A$ and $w'B$. This is given by:

$$\rho^2 = \frac{(\mathrm{cov}[v'A, w'B])^2}{\mathrm{var}[v'A]\,\mathrm{var}[w'B]} = \frac{(v'\Sigma_{AB}w)^2}{(v'\Sigma_{AA}v)(w'\Sigma_{BB}w)}$$

The maximum of ρ^2 in this case is known to be ρ_1^2, the largest eigenvalue of

$$\Sigma_{AA}^{-1}\Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA} \text{ or } \Sigma_{BB}^{-1}\Sigma_{BA}\Sigma_{AA}^{-1}\Sigma_{AB}$$

The maximum occurs when v is the eigenvector of

$$\Sigma_{AA}^{-1}\Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}$$

corresponding to ρ_1^2, and w is the eigenvector of

$$\Sigma_{BB}^{-1}\Sigma_{BA}\Sigma_{AA}^{-1}\Sigma_{AB}$$

corresponding to ρ_1^2. The positive square root $\sqrt{\rho_1^2}$ is called the first canonical correlation between A and B.

3.6 Krippendorff's Alpha[2]

Krippendorff's alpha is a statistical measure of the agreement achieved when categorizing a set of units of analysis in terms of the values of a variable. It is applicable to any number of coders, each assigning one value to any number of individual units of analysis. It adjusts to small sample sizes and handles missing data. The computed values of reliability are comparable across any number of coders, values, different metrics, and unequal sample sizes.

Although Krippendorff's alpha is normally used to measure the agreement among different raters, in our data, we do not have replication at the level of individual units. Each summary is graded by one and only one human assessor. However, in Chapter 6, Krippendorff's alpha will be used to measure the agreement of each rater's coding with him/herself. The statistic is calculated as follows:

$$\alpha = 1 - \frac{D_o}{D_e} = 1 - \frac{\sum_{u=1}^{N} \frac{m_u}{n} D_u}{D_e}$$

where the disagreement

$$D_u = \frac{1}{m_u(m_u-1)} \sum_{i=1,i'\neq i}^{m} \delta(c_{iu}, c_{i'u})$$

is the average difference $\delta(c_{iu}c_{i'u})$ between two values c_{iu} and $c_{i'u}$ over all $m_u(m_u - 1)$ pairs of values possible within unit u - without reference to coders. The observed disagreement

$$D_o = \sum_{u=1}^{N} \frac{m_u}{n} D_u = \frac{1}{n} \sum_{u=1}^{N} \frac{1}{m_u - 1} \sum_{i=1,i'\neq i}^{m} \delta(c_{iu}, c_{i'u})$$

[2] Most of this section was taken directly from [74]

is the average over all unit-wise disagreements in the data. The expected disagreement

$$D_e = \frac{1}{n(n-1)} \sum_{u=1,u'=1}^{N} \sum_{i=1,i'=1}^{m} \delta(c_{iu}, c_{i'u'}), [(i,u) \neq (i',u')]$$

is the average difference between any two values c_{iu} and $c_{i'u}$ over all $n(n-1)$ pairs of values possible within the reliability data - without reference to coders or units. In effect, D_e is the disagreement that is expected when the values used by all coders are randomly assigned to the given set of units.

Since the data in these experiments is Likert/ordinal, we use the appropriate difference function for ranks v and v' :

$$\delta_{ordinal}(v, v') = \left(\sum_{g=v}^{g=v'} n_g - \frac{n_v + n_{v'}}{2} \right)^2 .$$

The difference function δ reflects the metric properties of their variable. In general, all difference functions satisfy:

- $\delta(v, v') \geq 0$
- $\delta(v, v) = 0$
- $\delta(v, v') = \delta(v', v)$

3.7 Power of a Statistical Test

There are two main reasons practitioners are interested in the power of a test. The first reason is that a power calculation can be used to determine the sample size needed to conduct an experiment. The second reason is that we want to reject the null hypothesis whenever it is false, and a test with more power is more likely to do this. The following parts of this section explain the second reason further.

3.7.1 Tests with More Power

As will be mentioned in Chapter 4, when comparing statistical tests, we prefer the test with the most power. Power is a measure of the likelihood of a test to reject the null hypothesis when it is false. In order to investigate the change in power within a particular one parameter family of alternatives, I created a simulation to compare the power of several tests (signed rank test, paired t, and unpaired t). An important thing to consider when measuring the power of a test is the class of alternatives one is interested in being able to detect. In the realm of summarization scores, there is one type of alternative we always want to detect (and reject the null hypothesis in this situation). This type of alternative is humans vs. machines. Within the data we're dealing with (TAC 2008-2011), it is an almost certainty that any human summarizer receives a vector of scores that is significantly better than the vector of scores received by any machine summarizer. This fact was demonstrated in Chapter 4, where we showed that all three types of tests rejected the null hypothesis every time a machine system was compared against a human. However, it may not always be true that humans significantly outperform machine summarizers across the board, a fact pointed out in [20]. More generally, we also want to reject any of the alternatives discussed in Section 3.4.4.

In order to parameterize a class of alternatives of interest, I first grouped all the machine scores from the first 51 columns of Figure 3.1, which contains overall responsiveness scores from TAC 2011 and created the aggregate machine probability distribution (referred to as p_M). I then aggregated the scores in columns A–H from the same Figure to form the human probability distribution (referred to as p_H). These two distributions, p_M and p_H, are derived from all the scores received by any machine or human, respectively, on any document set. Each one has five components, corresponding to $P(X_M = k)$ or $P(X_H = k)$ for $k = 1, 2, 3, 4, 5$. The goal is to study a one-parameter family of alternatives, and the one I chose to look at is

$$p_t = (1 - t)p_M + tp_H.$$

The t parameter in this formula blends the two probability distributions together to create one that is "in the direction of p_H from p_M." At the extreme values of t, we have $p_0 = p_M$ and $p_1 = p_H$. It is well known that not all sets of documents are equally difficult to summarize. Because of this, it may be that a certain portion of the document sets lend themselves to quality automatic summarization, while the rest do not. We can then think of the parameter t as the percentage of documents on which machine summarizers perform as well as humans. As this parameter tends toward 1, the mixture distribution looks just like the humans, since at 1, machines are able to summarize 100% of documents as well as humans. In addition, although the p_t distribution is stochastically greater than the p_M distribution, the difference is not as great as the one between p_M and p_H, so we will not expect to see the 100% rejection rate we saw in Chapter 4.

To simulate data for this experiment, I let the mixture parameter t vary between 0.0 and 0.8, with steps as small as 0.05. The value $t = 0$ was a special case and was handled differently. For instance, the number of replications for $t = 0$ was 50,000 while the replications for every other value of t was 10,000. For all values of t, I did the following experiment the specified number of times. I generated samples of data for p_M and p_t in a way I will describe below. With the two samples in hand, I then computed the three test statistics mentioned above (signed rank test, paired t and unpaired t). (Further details on each test statistic are below.)

After the experiment is complete, the data can be analyzed. When $t = 0$, I have 50,000 of each of the three test statistics. Each group of 50,000 can be used to find a 95% quantile cutoff for each specific statistic. Then, for each value of $t \neq 0$, I have 10,000 values of each of the three different test statistics. Using these 10,000 values, I count how many are above the 95% quantile cutoff computed at $t = 0$ and plot the percentage on the graph. If for $t = 0.05$, we have 1000 values of the signed rank test statistic greater than the 95% cutoff for that statistic, then we plot 1,000/10,000 = 0.1 as the y-value above $t = 0.05$.

This is done for each value of t and for each of the three test statistics. The graphs in Figures 3.2 and 3.3 show the curves generated by each test's rejection frequency at each value of t. All three tests show very similar rates of rejection at any particular value of t and the rejection rates of all three tests are monotonically increasing as t increases.

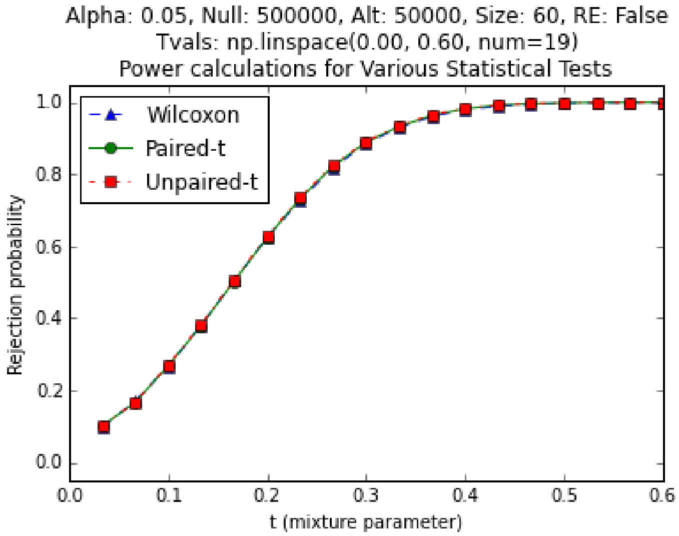


Figure 3.2: Power comparison without document random effects.

3.7.1.1 Test Statistics

At the lowest level of this simulation, we simply calculate three test statistics based on the samples X and Y. The three test statistics are the Wilcoxon signed rank, the paired t, and the unpaired t. All three statistics are normalized to account for the sample size. This is especially important with the signed rank test since we are following the Wilcoxon procedure and dropping the zeros/ties and normalizing by the new, reduced sample size. The paired and unpaired t-tests have a common numerator $(\bar{X} - \bar{Y})$, but their denominators are S_{X-Y} for the paired and $\sqrt{S_X^2/n_1 + S_Y^2/n_2}$ for the unpaired. Equal variances are not assumed in either t-test.

3.7.1.2 Sampling Techniques

For a particular value of t, we can create p_t to be compared against p_M. Given these two distributions, we can generate a sample of size 60 (the size chosen earlier) using statistical software. The results of this experiment are in Figure 3.2. The power curves line up perfectly here. The disadvantage of this method of sampling is that it does not create X and Y vectors that look like paired data, so it does not match up with the

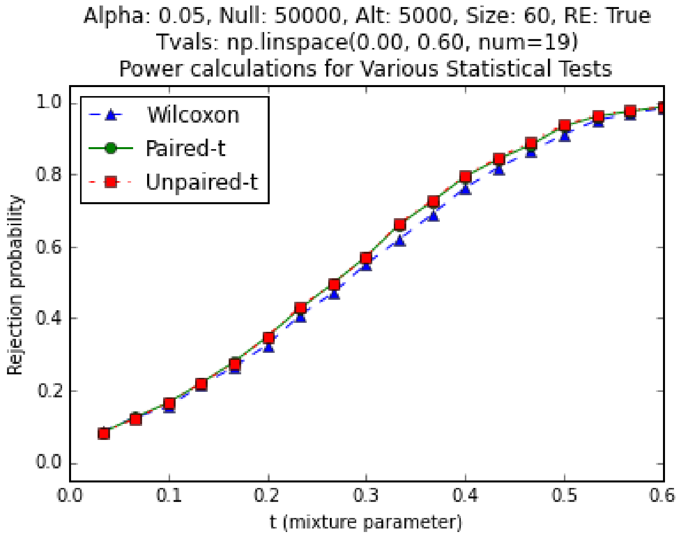


Figure 3.3: Power comparison with document random effects.

data we have. In order to account for this, I made use of the following method to create data that is paired.

3.7.1.3 Document Random Effects

The first step in creating paired data was to create a probability distribution for each document set. I followed the same procedure used for creating p_M and p_H and created p_d for $d = 1, \dots, 44$. This $\{p_d\}$ forms a sequence of probability vectors regarded as empirical probability distributions for Likert scores within individual document sets. In order to create two paired samples of size 60 (the size arbitrarily chosen earlier), I created a list of 60 document sets by re-sampling from the original list of 44. The result was a list of 60 values between 1 and 44, inclusive. To create the two vectors of 60 scores each, I then sampled two scores from each document set, one for p_M and one for p_t. The problem now, however, is that both of those scores come from the document set and don't differ according to p_M and p_t.

In order to sample from the document set's distribution p_d at the same time as from the machine's distribution p_M and the mixture distribution's p_t, I formulated conditional distributions $p_{M,d}$ for machine scores given document set d and machine-human mixture distributions $p_{t,d}$ for Likert scores given document set d. The idea behind this step is the following. Suppose $p_M = (0.1, 0.3, 0.3, 0.2, 0.1)$ and a particularly easy to summarize document set has $p_d = (0.0, 0.0, 0.0, 0.5, 0.5)$. The document set's summaries received scores of 4 and 5, each with 50% probability. Hence, when sampling from p_M and p_d simultaneously, I'd like the possible scores to only be 4s and 5s, since this document set never received a score lower than 4. In addition, since the machines' distribution indicates they are twice as likely to receive 4s than 5s, this should also factor into the scores to be sampled.

Here is one way to create a distribution that reflects some of the thoughts described above. Define $p_{M,d}$, the conditional distribution for machine scores given document set d as

$$p_{M,d}(k) = \frac{1}{p_M \cdot p_d} p_M(k) p_d(k),$$

where $p_M \cdot p_d$ denotes the usual dot product. In the same way,

$$p_{t,d}(k) = \frac{1}{p_t \cdot p_d} p_t(k) p_d(k),$$

and is interpreted as the conditional distribution of a Likert score given document set d when the score generating mechanism is a mixture of machine and human systems. In our example from above, $p_{M,d}$ would come out to be $(0.0, 0.0, 0.0, 0.667, 0.333)$. This reflects both the p_d distribution since it will only yield 4s and 5s and the p_M distribution since it yields 4s twice as often as 5s.

We now use the distributions $p_{M,d}$ and $p_{t,d}$ to sample a single score for X and a single score for Y. In addition, we use the same $p_{M,d}$ and $p_{t,d}$ to generate scores for this particular document d and this value of t. The goal in all of this was to sample scores in a way that reflects the reality of scores not being independent across document sets. By using these random effects, the simulated data ends up being paired in a way very similar to our actual data. If a document set is easy to summarize, it will effect both p_M and p_t and be more likely to yield higher scores for both. The end result is that the only mechanism causing a dependence between scores X_i, Y_i arising from the same document set in this simulation is the identity of the i^{th} document d itself and the fact that the conditional distributions given d were used to generate the scores. The scores are actually conditionally independent given document d but are unconditionally dependent.

3.7.2 Results

The curves in Figure 3.2 were generated under the simplifying assumption of identically distributed document sets. In this plot, there is really no discernible difference between any of the tests, and this is perhaps due the simplistic way in which the data was generated. The curves in Figure 3.3, however, trace the rejection probabilities of the same tests but in the new environment reminiscent of paired testing. The curves are still not very different, which is somewhat surprising considering the attempts to make the simulation reflect paired testing. It could be that I was not successful in doing this and that another format of testing would have demonstrated a higher power for the two paired tests used.

Chapter 4

Ranking Human and Machine Summarization Systems[1]

In the Text Analysis Conference's summarization task, the summarizers are ranked according to each human measure (one at a time) by averaging their scores over all document sets. In this chapter, the statistical appropriateness of this method is investigated and some alternatives are proposed that better distinguish between human and machine summarization systems. In particular, pairwise hypothesis testing is evaluated against the unpaired version and their powers are compared.

4.1 Background

This sort of ranking would be appropriate under certain statistical assumptions, such as one where the distributions from which each sample is drawn only differ by location shifts [54]. However, when a Kruskal-Wallis test (Section 3.4) of equal distributions is performed on this data, the resulting p-value is less than 10^{-12}. (This test was assuming normal continuous data, which is not the case here, so this p-value needs to be discounted. If the test were done along with a bootstrap simulation, a much

[1] This chapter's contents first appeared in [55].

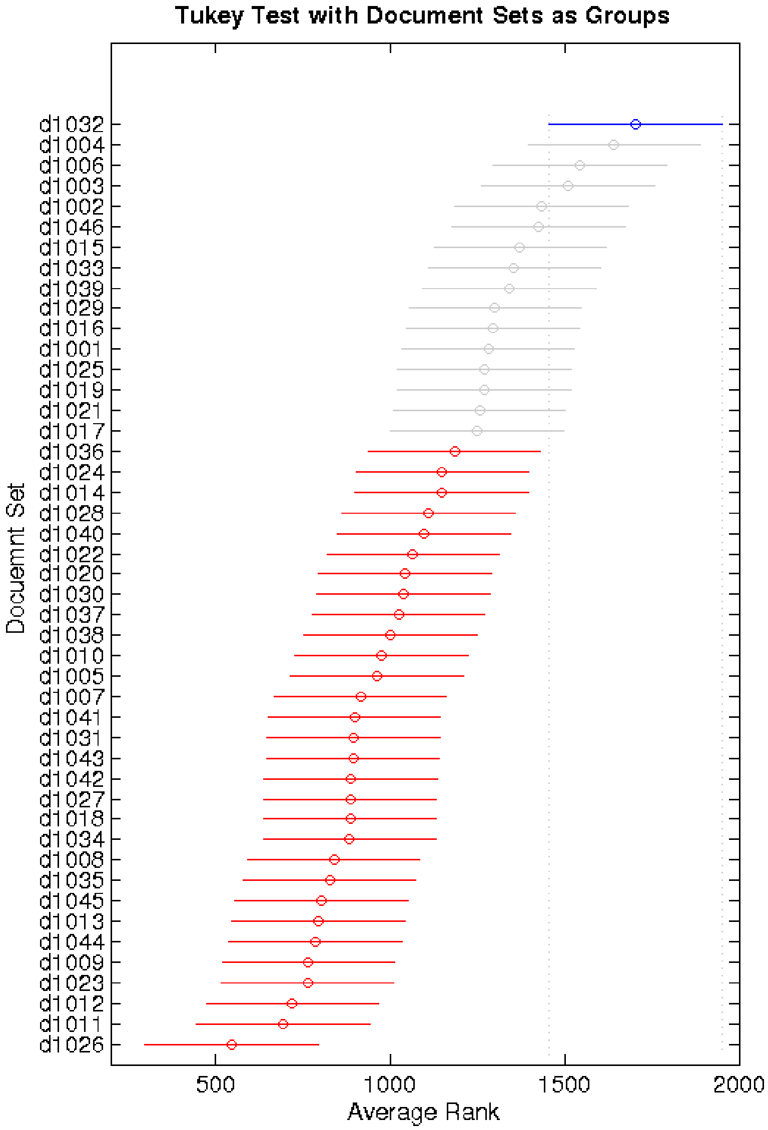


Figure 4.1: Confidence Intervals from a non-parametric Tukey's honest significant difference test for 46 TAC 2010 update document sets. The blue confidence interval (for document set d1032) does not overlap any of the 30 red intervals. Hence, the test concludes that 30 document sets have means significantly different from the mean of d1032. The confidence intervals were obtained from Matlab's multcompare function, which computes standard error using large sample normal theory of the studentized range distribution.

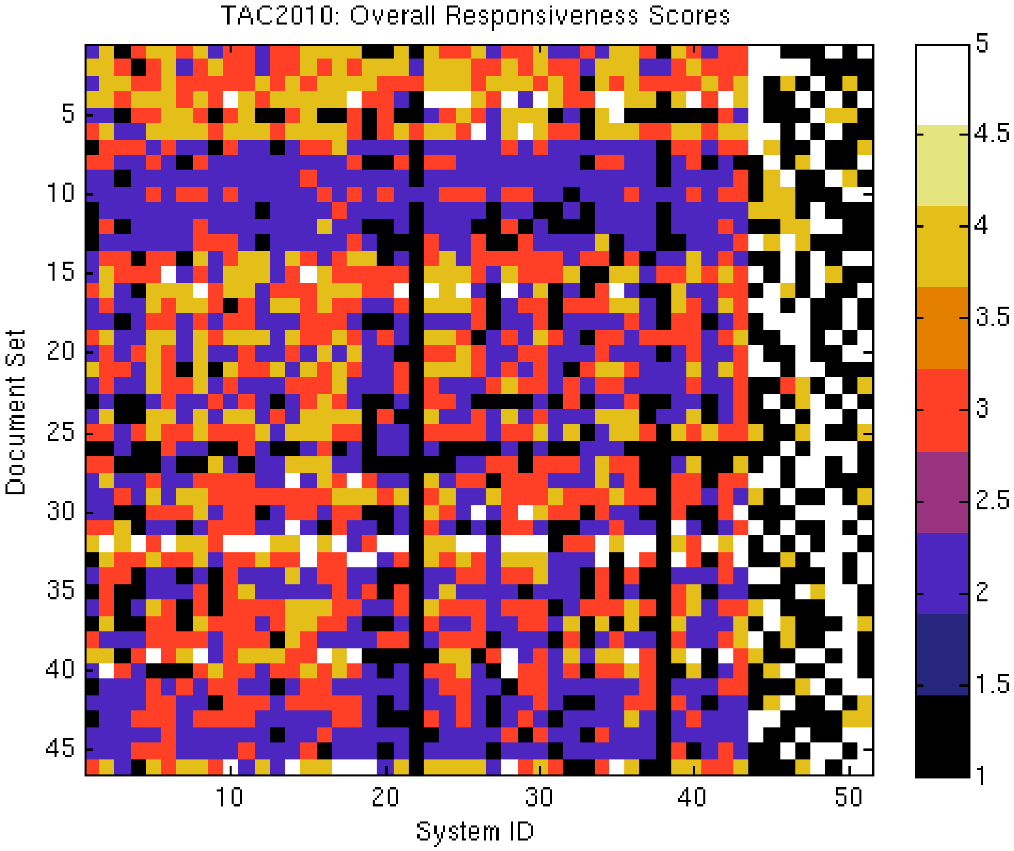


Figure 4.2: Overall responsiveness scores for the TAC 2010 update summary task, organized by document set (y-axis) and summarizer (x-axis). The 51 summarizers fall into two distinct groups: machine systems (first 43 columns) and humans (last 8 columns). Note that each human only summarized half of the document sets, thus creating 23 missing values in each of the last 8 columns. Black is used to indicate missing values in the last 8 columns and low scores in the first 43 columns.

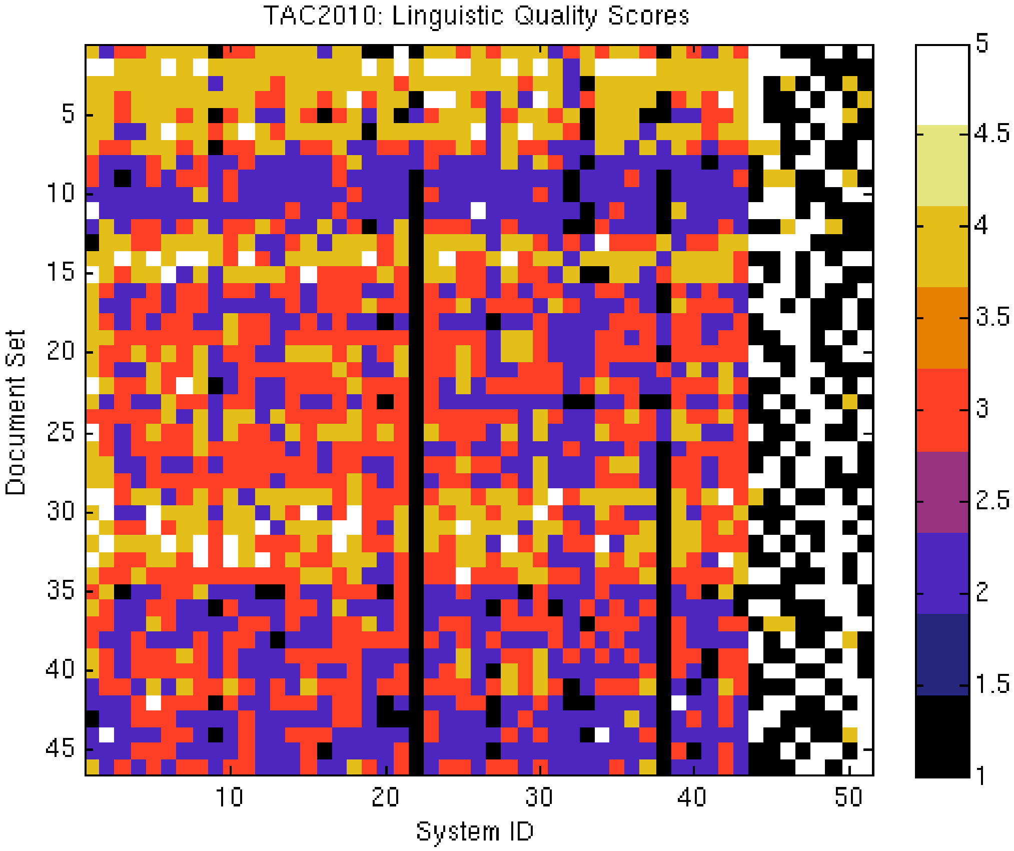


Figure 4.3: Linguistic (readability) scores for the TAC 2010 update summary task, organized by document set (y-axis) and summarizer (x-axis). The 51 summarizers fall into two distinct groups: machine systems (first 43 columns) and humans (last 8 columns). Note that each human only summarized half of the document sets, thus creating 23 missing values in each of the last 8 columns. Black is used to indicate missing values in the last 8 columns and low scores in the first 43 columns.

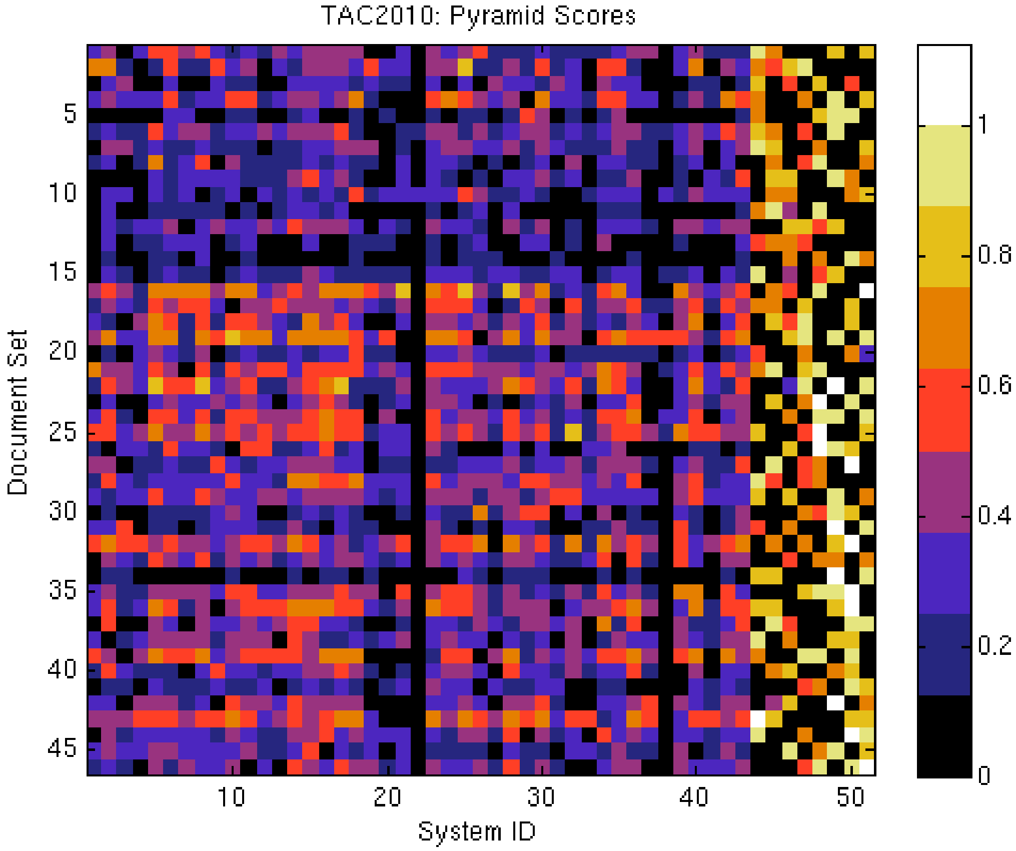


Figure 4.4: Pyramid scores for the TAC 2010 update summary task, organized by document set (y-axis) and summarizer (x-axis). The 51 summarizers fall into two distinct groups: machine systems (first 43 columns) and humans (last 8 columns). Note that each human only summarized half of the document sets, thus creating 23 missing values in each of the last 8 columns. Black is used to indicate missing values in the last 8 columns and low scores in the first 43 columns.

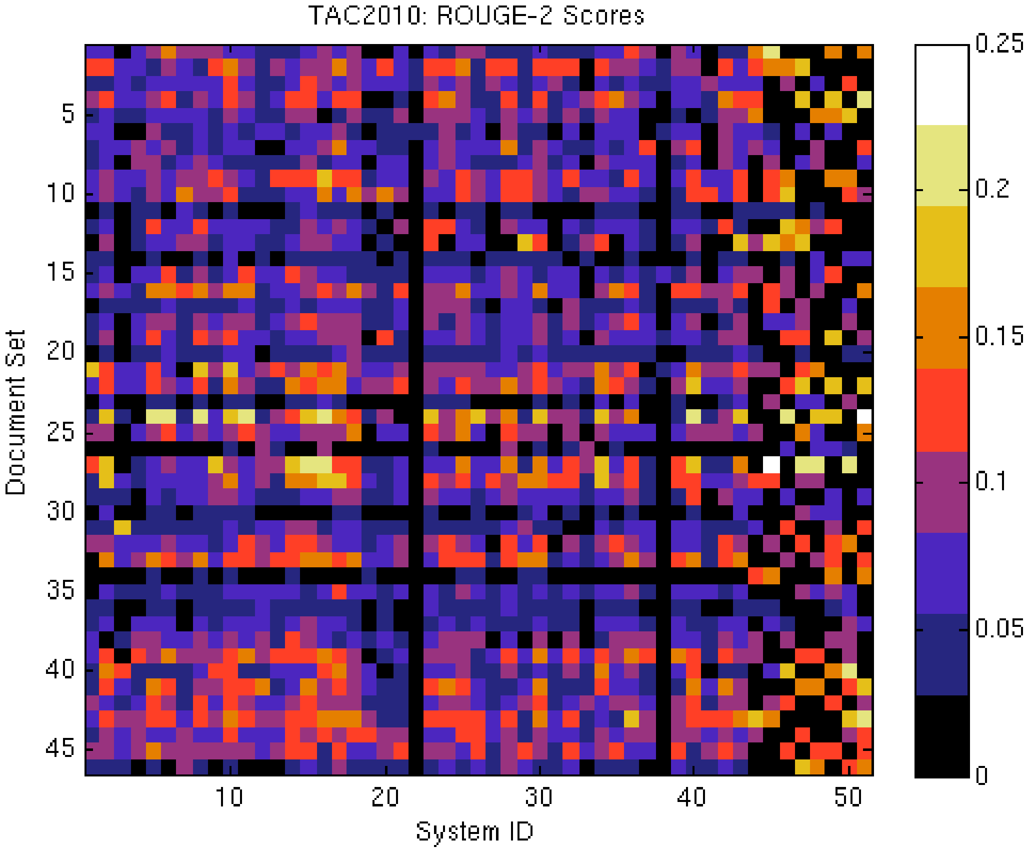


Figure 4.5: ROUGE-2 scores for the TAC 2010 update summary task, organized by document set (y-axis) and summarizer (x-axis). The 51 summarizers fall into two distinct groups: machine systems (first 43 columns) and humans (last 8 columns). Note that each human only summarized half of the document sets, thus creating 23 missing values in each of the last 8 columns. Black is used to indicate missing values in the last 8 columns and low scores in the first 43 columns.

more accurate p-value could be obtained, and probably would have also led to a highly significant rejection of the null hypothesis.) This provides evidence that a summary's score is not independent of the document set, and the effect can be seen in Figure 4.1. This figure shows the confidence bands computed by Tukey's honest significant difference test for each document set's difficulty (as measured by the mean rank responsiveness score for TAC 2010). The conclusion of the test is that average summarizer performance varies greatly across different document sets.

The difference in document sets is further illustrated in Figures 4.2 – 4.5, which show the scores of all summarizers on all the different document sets using standard human and automatic evaluation methods [9], using colors to indicate scores. These plots partially explain the result of the Kruskal-Wallis test. Some rows are clearly darker, indicating overall lower scores for the summaries of these documents, and the variances of the scores differ row-by-row. Hence, it may be advantageous to measure summarizer quality by accounting for the heterogeneity of documents within each test set. A non-parametric paired test like the Wilcoxon signed-rank is one way to do this, and the paired t-test is another.

In a 2008 COLING paper [5], Conroy and Dang noted an inconsistency in the best automatic evaluator at the time (ROUGE). When measured by human-produced scores, there is a significant gap in performance between machine systems and human summarizers. However, when measured by ROUGE, the gap is not present. In particular, in the DUC 2005-2007 data, some systems have ROUGE performance within the 95% confidence intervals of several human summarizers, but according to their pyramid, linguistic, and responsiveness scores, these same systems were clearly outside the intervals. Thus, the inexpensive automatic metrics at the time could not necessarily discern a difference between human summarizers and machine systems.

In this chapter we explore the use of document-paired testing for summarizer comparison. Our main approach is to consider each pair of two summarizers' sets of scores (over all documents) as a balanced two-sample dataset, and to assess that pair's mean difference in scores through a two-sample (paired or unpaired) t (Section 3.3) or Wilcoxon test (Section 3.2). Our goal has been to investigate whether human summarizer scores are uniformly different and better on average than machine summarizer scores, and to rate the quality of the statistical method (T or W, paired or unpaired) by the consistency with which the human versus machine scores show superior human performance. Our hope is that paired testing, using either the standard paired two-sample t-test or the distribution-free Wilcoxon signed-rank test, can provide greater power in the statistical analysis of automatic metrics such as ROUGE.

4.2 Size and Power of Tests

Statistical tests are generally compared by choosing rejection thresholds to achieve a certain small probability of Type I error (usually as $\alpha = .05$). Given multiple tests with the same Type I error, one prefers the test with the smallest probability of Type II error. Since power is defined to be one minus the Type II error probability, we prefer the test with the most power. Recall that a *test statistic* S depending on available data samples gives rise to a *rejection region* by defining rejection of the null hypothesis H_0 as the event $\{S \geq c\}$ for a *cutoff* or *rejection threshold* c chosen so that

$$P(S \geq c) \leq \alpha$$

for all probability laws compatible with the null hypothesis where the (nominal) *significance level* α is chosen in advance by the statistician, usually as $\alpha = .05$. However, in many settings, the null hypothesis comprises many possible probability laws, as here where the null hypothesis is that the underlying probability laws for the score samples of two separate summarizers are equal, without specifying exactly what that probability distribution is. In this case, the significance level may be an upper bound for the attained *size* of the test, defined as $\sup_{P \in H_0} P(S \geq c)$, the largest rejection probability $P(S \geq c)$ achieved by any probability law compatible with the null hypothesis. The power of the test then depends on the specific probability law Q from the considered alternatives in H_A. For each such Q, and given a threshold c, the power for the test at Q is the rejection probability $Q(S \geq c)$. These definitions reflect the fact that the null and alternative hypotheses are *composite*, that is, each consists of multiple probability laws for the data. One of the advantages of considering a *distribution-free*[2] two-sample test statistic such as the Wilcoxon is that the probability distribution for the statistic S is then the same for all (continuous, or non-discrete) probability laws $P \in H_0$, so that one cutoff c serves for all of H_0 with all rejection probabilities equal to α.

Two test statistics, say S and $\tilde{S}$, are generally compared in terms of their powers at fixed alternatives Q in the alternative hypothesis H_A, when their respective thresholds c, c^* have been defined so that the sizes of the respective tests, $\sup_{P \in H_0} P(S \geq c)$ and $\sup_{P \in H_0} P(\tilde{S} \geq c^*)$, are approximately equal. In this chapter, the test statistics under consideration are – in one-sided testing — the (unpaired) two-sample t test with pooled sample variance (T), the paired two-sample t test (T^p), and the (paired) signed-rank Wilcoxon test (W); and for two-sided testing, S is defined by the absolute value of one of these statistics. The thresholds c for the tests can be defined either by theoretical distributions, by large-sample approximations, or by data resampling (*bootstrap*) techniques, and (only) in the last case are these thresholds data dependent, or random. We explain these notions with respect to the two-sample data structure in which the scores from the first summarizer are denoted $X_1, \ldots, X_n$, where n is the number of documents with non-missing scores for both summarizers, and the scores from the second summarizer are $Y_1, \ldots, Y_n$. Let $Z_k = X_k - Y_k$ denote the document-wise differences between the summarizers' scores, and $\bar{Z} = \frac{1}{n} \sum_{k=1}^{n} Z_k$ be their average. Then the paired statistics are defined as

$$T^p \;=\; \sqrt{n(n-1)}\, \bar{Z} / (\sum_{k=1}^{n} (Z_k - \bar{Z})^2)^{1/2}$$

and

$$W \;=\; \sum_{k=1}^{n} \mathrm{sgn}(Z_k)\, R_k^+$$

[2] The Wilcoxon test is not distribution-free for discrete data. However, the discrete TAC data can be thought of as rounded continuous data, rather than as truly discrete data.

where $R_k^+ = \sum_{j=1}^n I(|Z_j| \leq |Z_k|)$ is the rank of $|Z_k|$ among $|Z_1|, \ldots, |Z_n|$. Note that under both null and alternative hypotheses, the variates Z_k are assumed to be independent and identically distributed (*i.i.d.*), while under H_0, the random variables Z_k are symmetric about 0.

The t-statistic T^p is 'parametric' in the sense that exact theoretical calculations of probabilities $P(a < T^p < b)$ depend on the assumption of normality of the differences Z_k, and when that holds, the two-sided cutoff $c = c(T^p)$ is defined as the $1-\alpha/2$ quantile of the t_{n-1} distribution with $n-1$ degrees of freedom. However, when n is moderately or very large, the cutoff is well approximated by the standard-normal $1-\alpha/2$ quantile $z_{\alpha/2}$, and T^p becomes approximately nonparametrically valid with this cutoff, by the Central Limit Theorem. The Wilcoxon signed-rank statistic W has theoretical cutoff $c = c(W)$ which depends only on n, whenever the data Z_k are continuously distributed; but for large n, the cutoff is given simply as $\sqrt{n^3/12} \cdot z_{\alpha/2}$. When there are ties (as might be common in discrete data), the calculation of cutoffs and p-values for Wilcoxon becomes slightly more complicated and is no longer fully nonparametric except in a large-sample approximate sense.

The situation for the two-sample unpaired t-statistic T currently used in TAC evaluation is not so neat. Even when the two samples $\mathbf{X} = \{X_k\}_{k=1}^n$ and $\mathbf{Y} = \{Y_k\}_{k=1}^n$ are independent, exact theoretical distribution of cutoffs is known only under the parametric assumption that the scores are normally distributed (and in the case of the pooled-sample-variance statistic, that $\text{Var}(X_k) = \text{Var}(Y_k)$.) However, an essential element of the summarization data is the heterogeneity of documents. This means that while $\{X_k\}_{k=1}^n$ can be viewed as *i.i.d.* scores when documents are selected randomly – and not necessarily equiprobably – from the ensemble of all possible documents, the Y_k and X_k samples are *dependent.* Still, the pairs $\{(X_k, Y_k)\}_{k=1}^n$, and therefore the differences $\{Z_k\}_{k=1}^n$, are *i.i.d.* which is what makes paired testing valid. However, there is no theoretical distribution for T from which to calculate valid quantiles c for cutoffs, and therefore the use of the unpaired t-statistic cannot be recommended for TAC evaluation.

Even though there is not always a theoretical distribution for our test statistic, there are ways to ascertain the approximate validity of theoretically derived large-sample cutoffs for test statistics. In the age of plentiful and fast computers, we can employ the powerful computational machinery of the *bootstrap* [13].

The idea of bootstrap hypothesis testing [13], [2] is to randomly sample with replacement (the rows with non-missing data in) the dataset $\{(X_k, Y_k)\}_{k=1}^n$ in such a way as to generate representative data that plausibly *would* have been seen if two-sample score data had been generated from two equally effective summarizers with score distributional characteristics like the pooled scores from the two observed summarizers. We have done this in two distinct ways, each creating 2000 datasets with n paired scores:

MC *Monte Carlo Method.* For each of many iterations (in our case 2000), define a new dataset $\{(X'_k, Y'_k)\}_{k=1}^n$ by independently swapping X_k and Y_k with probability 1/2. Hence, $(X'_k, Y'_k) = (X_k, Y_k)$ with probability 1/2 and (Y_k, X_k) with probability 1/2.

HB *Hybrid MC/Bootstrap.* For each of 2000 iterations, create a re-sampled dataset $\{(X''_k, Y''_k)\}_{k=1}^n$ in the following way. First, sample n pairs (X_k, Y_k) with replacement from the original dataset. Then, as above, randomly swap the components of each pair, each with 1/2 probability.

Both of these two methods can be seen to generate two-sample data satisfying H_0, with each score sample's distribution obtained as a mixture of the distributions actually generating the $\mathbf{X}$ and $\mathbf{Y}$ samples. The *empirical q^{th} quantiles* for a statistic $S = S(\mathbf{X}, \mathbf{Y})$ such as $|W|$ or $|T^p|$ are estimated from the resampled data as $\hat{F}_S^{-1}(q)$, where $\hat{F}_S(t)$ is simply the fraction of times (out of 2000) that the statistic S applied to the constructed dataset had a value less than or equal to t. The upshot is that the $1-\alpha$ empirical quantile for S based on either of these simulation methods serves as a data dependent cutoff c attaining approximate size α for all H_0-generated data. The MC and HB methods will be employed in Section 4.4 to check the theoretical p-values.

It is worth noting that these two methods are similar to Fisher's permutation test, which considers every possible re-labeling of the data and derives the significance of the test statistic from the range of values it can achieve in other possible arrangements. The permutation test is often computationally infeasible, even given moderately small sizes of data, and can be approximated by the above procedures.

4.3 Relative Efficiency of W versus T^p

Statistical theory does have something to say about the comparative powers of paired W versus T^p statistics. These statistics have been studied [54], in terms of their *asymptotic relative efficiency* for location-shift alternatives based on symmetric densities ($f(z - \vartheta)$ is a location-shift of $f(z)$). For many pairs of parametric and rank-based statistics $S, \tilde{S}$, including W and T^p, the following assertion has been proved for testing H_0 at significance level α.

First assume the Z_k are distributed according to some density $f(z - \vartheta)$, where $f(z)$ is a symmetric function ($f(-z) = f(z)$). Next assume $\vartheta = 0$ under H_0. When n gets large the powers at any alternatives with very small $\vartheta = \gamma/\sqrt{n}$, $\gamma \neq 0$, can be made asymptotically equal by using samples of size n with statistic S and of size $\rho \cdot n$ with statistic $\tilde{S}$. Here $\rho = ARE(S, \tilde{S})$ is a constant not depending on n or γ but definitely depending on f, called the *asymptotic relative efficiency* of S with respect to $\tilde{S}$. (The smaller $\rho < 1$ is, the more statistic $\tilde{S}$ is preferred among the two.)

Using this definition, it is known (Randles and Wolfe 1979 [54], Sec. 5.4 leading up to Table 5.4.7 on p. 167) that the Wilcoxon signed-rank statistic W provides greater robustness and often much greater efficiency than the paired T, with ARE which is 0.95 with f a standard normal density, and which is never less than 0.864 for any symmetric density f. However, in our context, continuous scores such as pyramid exhibit document-specific score differences between summarizers which often have approximately normal-looking histograms, and although the alternatives perhaps cannot be viewed as pure location shifts, it is unsurprising in view of the ARE theory cited above that the W and

Table 4.1: Number of significant differences found when testing for the difference of all pairs of summarization systems (including humans).

	2008: 2145 = $\binom{66}{2}$ pairs			2009: 1830 = $\binom{61}{2}$ pairs			2010: 1275 = $\binom{51}{2}$ pairs		
Metric	**Unpair-T**	**Pair-T**	**Wilc.**	**Unpair-T**	**Pair-T**	**Wilc.**	**Unpair-T**	**Pair-T**	**Wilc.**
Linguistic	1234	**1416**	1410	1000	**1182**	1173	841	**939**	934
Overall	1202	**1353**	1342	982	**1149**	1146	845	**894**	889
Pyramid	1263	1417	**1418**	1075	**1238**	1216	875	**933**	926
ROUGE-2	1243	1453	**1459**	1016	1182	**1193**	812	938	**939**
ROUGE-SU4	1333	1493	**1507**	1059	1241	**1254**	894	**983**	976

Table 4.2: Number of significant differences resulting from $8 \times (N - 8)$ tests for human-machine system means or signed-rank comparisons.

	2008: $464 = 58 \times 8$ pairs			2009: $424 = 53 \times 8$ pairs			2010: $344 = 43 \times 8$ pairs		
Metric	**Unpair-T**	**Pair-T**	**Wilc.**	**Unpair-T**	**Pair-T**	**Wilc.**	**Unpair-T**	**Pair-T**	**Wilc.**
Linguistic	464	464	464	424	424	424	344	344	344
Overall	464	464	464	424	424	424	344	344	344
Pyramid	464	464	464	424	424	424	344	344	344
ROUGE-2	375	**409**	402	323	**350**	341	275	**309**	305
ROUGE-SU4	391	**418**	414	354	**378**	373	324	**331**	328

T paired tests have very similar performance. Nevertheless, as we found by statistical analysis of the TAC data, both are far superior to the unpaired T-statistic, with either theoretical or empirical bootstrapped p-values.

4.4 Testing Setup and Results

To evaluate our ideas, we used the TAC data from 2008-2010 and focused on three manual metrics (overall responsiveness, pyramid score, and linguistic quality score) and two automatic metrics (ROUGE-2 and ROUGE-SU4). We make the assumption, backed by both the scores given and comments made by NIST summary assessors [3], that automatic summarization systems do not perform at the human level of performance. As such, if a statistic based on an automatic metric, such as ROUGE-2, were to show fewer systems performing at human level of performance than the statistic of averaging scores, such a statistic would be preferable because of its greater power in the machine vs. human

[3] Assessors have commented privately at the Text Analysis Conference 2008, that while the origin of the summary is hidden from them, "we know which ones are machine generated." Thus, automatic summarization fails the Turing test of machine intelligence [71]. This belief is also supported by [5] and [6]. Finally, our own results show no matter how you compare human and machine scores all machine systems score significantly worse than humans.

summarization domain.

For each of these metrics, we first created a score matrix whose (i, j)-entry represents the score for summarizer j on document set i (these matrices generated the color plots in Figures 4.2 – 4.5). We then performed a Wilcoxon signed-rank test on certain pairs of columns of this matrix (any pair consisting of one machine system and one human summarizer). As a baseline, we did the same testing with a paired and an unpaired t-test. Each of these tests was based on large sample normal theory approximations and resulted in a p-value based on that theory. We counted how many p-values were less than .05 and called these the significant differences.

The results of these tests (shown in Table 4.2), were somewhat surprising. Although we expected the nonparametric signed-rank test to perform better than an unpaired t-test, we were surprised to see that a paired t-test performed even better. All three tests always reject the null hypotheses when human metrics are used. This is what we would like to happen with automatic metrics as well. As seen from the table, the paired t-test and Wilcoxon signed-rank test offer a good improvement over the unpaired t-test.

The results in Table 4.1 are less clear, but still positive. In this case, we are comparing pairs of machine summarization systems. In contrast to the human vs. machine case, we do not know the truth here. However, since the number of significant differences increases with paired testing here as well, we believe this also reflects the greater discriminatory power of paired testing. We also note that this result contradicts the behavior observed by Smucker, Allan, and Carterette [64], whose experiment showed the Wilcoxon signed rank test to have "poor ability to detect significance" and "potential to lead to false detections of significance." At this time, we do not know the cause of this discrepancy.

We now apply the Monte Carlo and Hybrid Monte Carlo to check the theoretical p-values reported in Tables 4.1 and 4.2. The empirical quantiles found by these methods generally confirm the theoretical p-value test results reported there, especially in Table 4.2. In the overall tallies of all comparisons (Table 4.1), it seems that the bootstrap results (comparing only W and the un-paired T) make W look still stronger for linguistic and overall responsiveness versus the T; but for the pyramid and ROUGE scores, the bootstrap p-values bring T slightly closer to W although it still remains clearly inferior, achieving roughly 10% fewer rejections.

4.5 Conclusions and Future Work

In this chapter we observed that summarization systems' performance varied significantly across document sets on the Text Analysis Conference (TAC) data. This variance in performance suggested that paired testing may be more appropriate than the t-test currently employed at TAC to compare the performance of summarization systems. We proposed a non-parametric test, the Wilcoxon signed-rank test, as a robust more powerful alternative to the t-test. We estimated the statistical power of the t-test and the Wilcoxon signed-rank test by calculating the number of machine systems whose performance was significantly different than that of human summarizers. Since human assessors score machine systems as not achieving human performance in either content or responsiveness, automatic metrics such as ROUGE should ideally indicate this distinction. We found that the paired Wilcoxon test significantly increases the number of machine systems that score significantly different than humans when the pairwise test is performed on ROUGE-2 and ROUGE-SU4 scores. Thus, we demonstrated that the Wilcoxon paired test shows more statistical power than the t-test for comparing summarization systems.

Consequently, the use of paired testing should not only be used in formal evaluations such as TAC, but also should be employed by summarization developers to more accurately assess whether changes to an automatic system give rise to improved performance.

Further study is needed to analyze more summarization metrics such as those proposed at the recent NIST evaluation of automatic metrics, Automatically Evaluating Summaries of Peers (AESOP) [68]. As metrics become more sophisticated and aim to more accurately predict human judgments such as overall responsiveness and linguistic quality, paired testing seems likely to be a more powerful statistical procedure than the unpaired t-test for head-to-head summarizer comparisons.

Throughout our research for this chapter, we treated each separate kind of scores on a document set as data for one summarizer to be compared with the same kind of scores for other summarizers. However, it might be more fruitful to treat *all* the scores as multivariate data and compare the summarizers that way. Multivariate statistical techniques such as Principal Component Analysis may play a constructive role in suggesting highly discriminating new composite scores, perhaps leading to statistics with even more power to measure a summary's quality.

ROUGE was inspired by the success of the BLEU (BiLingual Evaluation Understudy), an n-gram based evaluation for machine translation [48]. It is likely that paired testing may also be appropriate for BLEU as well and will give additional discriminating power between machine translations and human translations. Paired testing was considered in the context of randomized significance tests, but the conclusion there was that it made very little difference in the experiment conducted [21].

Chapter 5

Better Metrics to Automatically Predict the Quality of a Text Summary[1]

In this chapter we demonstrate a family of algorithms for predicting the quality of a text summary. The inputs to the algorithm are features computed directly from the text, and include a combination of linguistic and content features. These features were combined using several types of linear models assessed by means of a cross-validation approach with a single data split between training and test datasets. The resulting algorithms achieve significantly better correlation with human judgments of summary quality than the previous standard for automatic text summarization evaluation, ROUGE.

5.1 Background

Due to the now common explosion of information, we are often faced with too much to read. Search engines have improved dramatically over the past several decades, so much so that they frequently return too many relevant documents. Tools that enable us to sift

[1] This chapter's contents first appeared in [57].

through data to find relevant documents and, more generally, information of the greatest interest to us, are much needed. One approach to deal with this overload of information is to produce summaries of the documents. The problem of single document summarization was first introduced over 55 years ago by Luhn [38]. Since then, hundreds of papers on text summarization have been published. In 1995 McKeown and Radev [39] introduced multi-document summarization where information from a collection of documents on the same topic is summarized. This approach naturally leads to a top-down navigation of documents to extract relevant information. A series of evaluations of summarization methods has been conducted over the last dozen years and the data from these will be described in Section 5.3. The most recent of these has been the Text Analysis Conference (TAC) [68] which is sponsored by the National Institute of Standards and Technology (NIST). Each year at TAC, several dozen summarization systems are evaluated by NIST based on several criteria.

The two main types of summarization evaluation are extrinsic and intrinsic. In extrinsic evaluation, summaries are evaluated according to how successfully an external task can be completed using only the summary. In intrinsic evaluation, a summary's score is derived from the summary itself, perhaps by comparison with a human-written, gold standard summary. For the purposes of this work, we focus only on intrinsic evaluation.

The ultimate intrinsic test of the quality of a summary is human judgment of its content, linguistic quality, and overall responsiveness to the needs of the given information task. These human judgments will be described in more detail in Section 5.3. Suffice it to say that such human-based evaluation, while absolutely necessary to evaluate the task, is very time consuming. Because of this, we seek to find automatic evaluation metrics which, as closely as possible, will correlate with human judgments of text summaries. In recent years, automatic summarization systems have made great gains in their performance. Today, top performing summarization systems outperform humans when measured by traditional automatic metrics, despite the fact that the measured performance, as judged by human evaluators, indicate that the automatic systems perform significantly worse than humans. Mathematically, such a metric gap is a discontinuity in the function relating the automatic and manual metrics and is illustrated in Figure 5.1. This metric gap will be discussed further in the next few paragraphs.

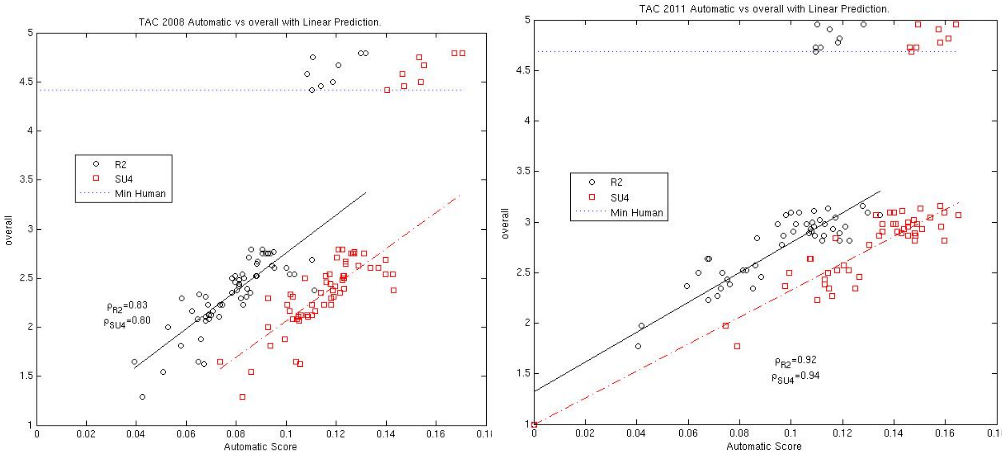


Figure 5.1: TAC 2008/2011 task results.

This plot shows ROUGE-2 and ROUGE-SU4, two of the standard baseline evaluation systems. The regression line only goes as far as the lowest-scoring human summary. In the ideal picture, the cluster of human summarizers at the top of each plot would be further to the right and directly along the prediction line that ROUGE fits with the machine summarizers.

Measuring correlation between automatic metrics and human judgments is a natural way to measure the performance of an automatic metric. When NIST recently evaluated automatic summarization metrics, the three widely used correlation metrics—Pearson, Spearman, and Kendall's tau—were used. Of these three metrics we present only Pearson, primarily for space considerations, and we direct the reader to [68] for results using Spearman and Kendall's tau. Secondly, we favor Pearson as a metric since, of the three, it penalizes metrics for discontinuities such as those illustrated in Figure 5.1.

The baseline summarization evaluation method to which many others are often compared is Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [34] and is a (word) n-gram-based approach for comparing one summary to one or more human-generated summaries (note that to evaluate a human summary, one simply compares it to the other human summaries). Generally, bigram-based versions of ROUGE give the best correlation with human judgments of a summary.

To illustrate the strengths and weaknesses of ROUGE as a measure of a summarization system's performance, we refer again to the two scatter plots in Figure 5.1. In these plots, ordered pairs of each system's average ROUGE scores and average "overall responsiveness" scores are given for each automatic summarization system as well as for 8 human summarizers. The plots illustrate two of the best ROUGE scoring approaches, ROUGE-2 which measures the bigram similarity of a summary against human-generated summaries and ROUGE-SU4 which measures bigram similarity but allows for a skip distance of up to 4 words. Both plots show 2 major groups of data which, not surprisingly, correspond to the machine-generated and human-generated summaries. In 2008, there was a wide gap in the performance in average overall responsiveness (the human judgment) while the best systems scored in the lower range of human performance for both ROUGE metrics. Three years later, we see that the best machine-generated summaries have made improvement in both ROUGE and responsiveness, so much so that some systems now exceed the performance of humans in the ROUGE metric. However, their performance in responsiveness pales in comparison with that of humans.

This inability of ROUGE to adequately predict human judgments of summaries gave rise to a "meta-evaluation", an evaluation of evaluation methods. The task is called AESOP or Automatically Evaluating Summaries of Peers and has been part of TAC for the last two years. In this chapter, we propose using both content-oriented features, similar in spirit to ROUGE, in conjunction with low-level linguistic features to produce a metric that correlates well between human-generated summaries and those produced by machine-generated summarization systems. The features were combined based on training on previous years of TAC data using linear algebraic and statistical methods. They were then used to predict scores for the 2011 data prior to their release. The resulting metrics more accurately predicted the current performance gap between human and machine generated summaries than ROUGE. In addition, we have applied our methods to the 2008 TAC data to further validate their ability to predict the human-machine performance gap.

The rest of this chapter is organized as follows: Section 5.2 covers related work in text summarization evaluation. We discuss the origin of our data in Section 5.3. In Section 5.4, we define the linguistic and content features that go into our supervised learning algorithms and in Section 5.5, we discuss how those algorithms select subsets of the features. Section 5.6 describes our results and Section 5.7 contains our conclusions and ideas for future work.

5.2 Related Work

Previous work has looked at extensions to (word) n-gram approaches for judging the quality of a summary relative to one or more human-generated summaries. Conroy and Dang [5] analyzed summarization evaluation data from the Document Understanding Conferences (DUC [12], the TAC predecessor) for the years 2005–2007. They proposed using robust regression to find canonical correlation coefficients to improve the correlation between the automatically-generated metric and human-generated metrics. Their approach used several variants of the ROUGE (word) n-gram scoring to produce a new metric, called ROUGE Optimal Summarization Evaluation (ROSE). ROSE was successful in improving the correlation within a given year, which would be useful for researchers producing new summarization methods to build an interpolation model to compare their new method with methods that were evaluated by humans. However, the method was not successful in improving the correlation when a model was built using one year's data, say 2005, and then applying it to another year's data, say 2006. The authors did demonstrate that if the actual linguistic quality scores were known, then such a cross-year improvement could be attained.

For "update summaries", the base summary (or summaries), which encapsulates the readers' knowledge of the topic so far, must be taken into account. An update summary is a text summary that is generated on an “evolving topic” and summaries are generated for new documents with a focus on what is novel in the new set of documents. Evaluation of such summaries can pose a challenge. An approach that uses the ROUGE content metrics comparing the information in the update summary to that of the previously generated summaries on the topic was proposed by Conroy, Schlesinger and O’Leary [7]. Nouveau-ROUGE computes a linear combination of the similarity of the summary with not only the current human-generated summaries on the topic but those corresponding to the previously known information. Such an approach was shown to improve correlation with both pyramid and overall responsiveness metrics, two important human-generated scores given to summaries.

Oliveira *et al.* [11] proposed a system called Valuation using Enhanced Rationale Technique (VERT-F). VERT-F compares the (word) bigrams of summaries to be evaluated with human-generated summaries by computing a statistic which is a combination of 4 others: the χ^2 statistic, and three metrics from information retrieval—precision, recall, and the geometric mean of these ($F-$measure).

Giannakopoulos *et al.* [18] proposed a novel (character) n-gram approach, AutoSummENG (AUTOmatic SUMMary Evaluation based on N-gram Graphs), which was designed to be language independent. The approach builds graphs for a given summary based on n-gram counts. Graph similarity metrics are then used to compare summaries. The resulting metrics were shown to be competitive with ROUGE scoring. (Giannakopoulos *et al.* [19] turned the evaluation around and used it as a method to generate summaries.)

The Merged Model Graph method (MeMoG) [17] is a variation on AutoSummENG where instead of comparing summaries to summaries, a summary’s graph is compared to a merged graph of 3 human-generated summaries. We will compare the performance of our approach to AutoSummENG, MeMoG, and VERT-F, as well as several variations of ROUGE in Section 5.6.

Pitler, Louis, and Nenkova [50] went in a different direction and sought to automatically evaluate the linguistic quality of a summary. They tested numerous high-level and low-level features such as cosine similarity between sentences, co-reference information, and number of syllables in a sentence's words in an attempt to separately predict several aspects of linguistic quality, including grammaticality, non-redundancy, referential clarity, focus, and structure and coherence. Their best results were 90% accuracy for pairwise comparisons of competing systems and 70% accuracy for ranking summaries. Kanungo and Orr [27] also attempted to estimate readability, but focused on web summaries. They used summary features to fit a gradient boosted decision tree, and reported an increase in correlation with editorial judgments.

A graph-based method is presented in [32], where the authors compute term weights to give more credit to essential words in reference sentences (*i.e.*, sentences from a human-generated summary) when used to automatically measure the quality of a human generated summary. This competitive approach uses a centrality score as well as sentence clustering to weight co-occurring words in a sentence. The modifications of their approach were developed after the TAC data were released and are a slight variation of what the authors prepared for TAC 2011. The results of this approach are given in Figures 5.2, 5.3, and 5.4 and are labeled 10, 13, 20 and 24.

The following two papers consider the challenge of evaluating machine-produced summaries without model human-written summaries. Steinberger and Ježek [66] propose using Latent Semantic Analysis (LSA) for summarization evaluation. The idea is that LSA can identify the most important topics in a reference text (full document or abstract) and that a summary can be judged based on how similar its topics are to those in the reference text. They produce similarity scores by computing angles between singular vectors and combinations of singular vectors and achieve good results with model summaries and separately with full text documents.

In [58], Saggion *et al.* expand on the work of Louis and Nenkova [36] by employing Jensen–Shannon divergence to contrast the probability distribution of a summary with that of its reference text. They show that substituting the full document for the model

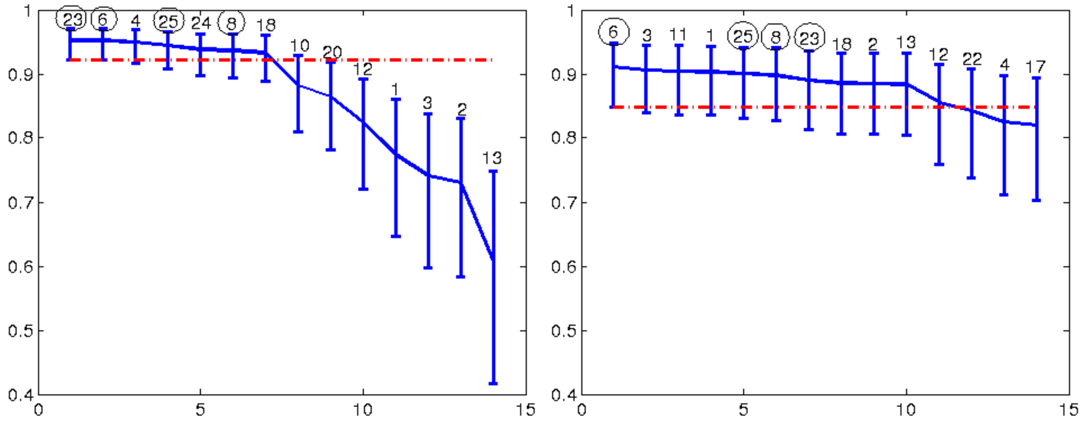


Figure 5.2: Pearson correlation with pyramid top 14 AESOP metrics: all peers and no models (2011).

Systems 1–3 are the baseline evaluation systems (see Section 5.6). Our systems are numbers 6, 8, 23 and 25. Systems 12 (AutoSummENG) and 18 (MeMoG) are due to Giannakopoulos *et al.* [18]. System 4 is VERT-F, due to Oliveira *et al* [11]. The vertical bar shown for each system is a 95% confidence interval (uses t-distribution and assumes normal data) for its Pearson correlation with the human metric listed at the top of the figure.

summary works almost as well in certain cases (generic and topic-based multi-document summarization) but performs much worse in others (summarization of biographical information and summarization of opinions in blogs).

We believe our work is the first to simultaneously combine the different components we used for the scoring of text summaries. We used a novelty metric, content features and various linguistic features to create a family of algorithms able to accurately predict the responsiveness and readability of a text summary.

5.3 The Data

The data for this evaluation is taken from the 2008–2011 update tasks from TAC. The summarization task consists of two sets of 10 documents in each of 40+ document sets. The first set of 10 is the “base”; the second set occurs later in time and is called the “update”. Participants’ programs create two summaries: One that summarizes the first ten documents and another that summarizes the second ten, focusing only on novel

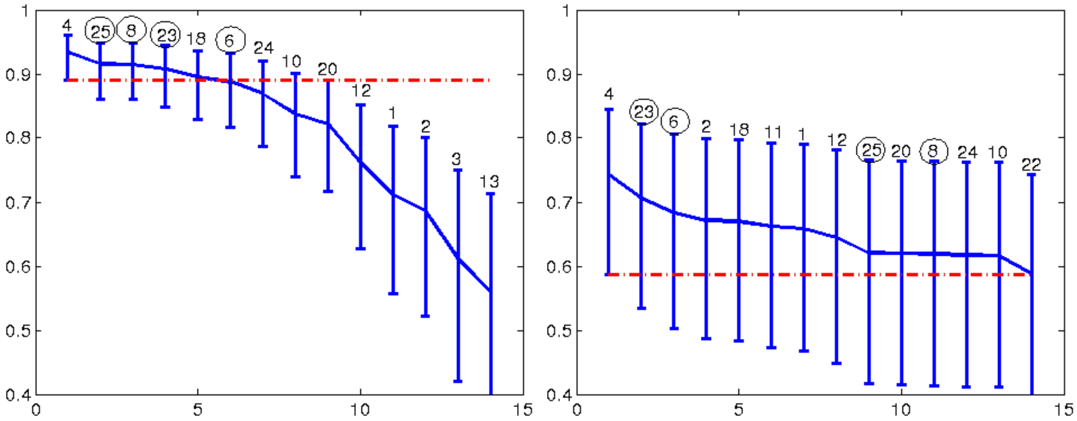


Figure 5.3: Pearson correlation with readability top 14 AESOP metrics: all peers and no models (2011).

Systems 1–3 are the baseline evaluation systems (see Section 5.6). Our systems are numbers 6, 8, 23 and 25. Systems 12 (AutoSummENG) and 18 (MeMoG) are due to Giannakopoulos *et al.* [18]. System 4 is VERT-F, due to Oliveira *et al* [11]. The vertical bar shown for each system is a 95% confidence interval for its Pearson correlation with the human metric listed at the top of the figure.

information not contained in the first ten.

In 2010, and continuing to 2011, the notion of a “guided summary” was added. Five categories were established and a set of “aspects,” specific information relevant to the category, was defined for each. Every data set was linked to one category. Inclusion of information that addressed the various aspects was a major focus of the evaluation for both the base and update summaries.

NIST has sponsored a summarization evaluation every year since 2001 (DUC through 2007; TAC from 2008 on). Methods to evaluate the submitted summaries have been a major effort and have, of course, evolved over the years. Beginning in 2004, ROUGE [34] became the automatic method used, replacing a measure of precision and recall of the sentences in the generated summaries. Due to the improved quality of generated summaries, ROUGE scores are no longer as good a predictor of summary quality (when compared with human evaluation of summaries) as they once were. Hence the search for new automatic tools to measure summary quality.

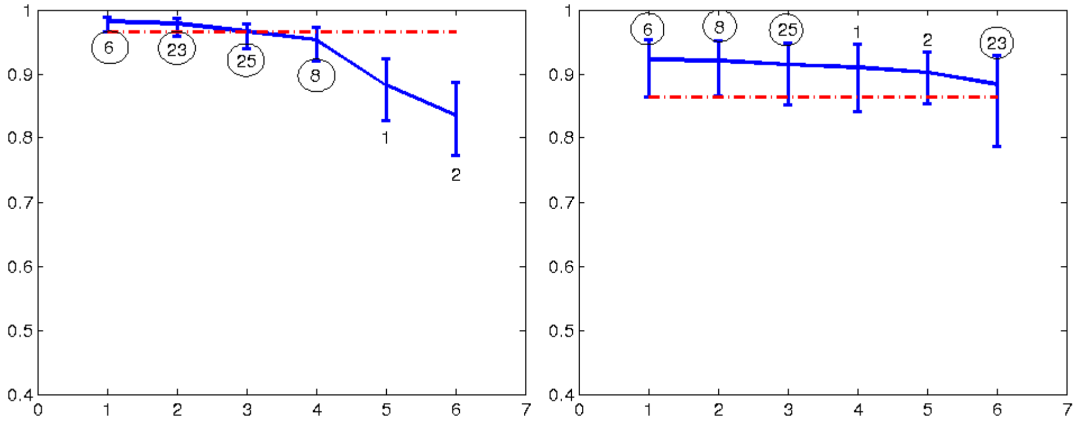


Figure 5.4: Pearson correlation with responsiveness all peers and no models (2008). Systems 1 and 2 are the baseline evaluation systems (ROUGE-2 and ROUGE-SU4, respectively). Our systems are numbers 6, 8, 23 and 25. The vertical bar shown for each system is a 95% confidence interval for its Pearson correlation with the human metric listed at the top of the figure.

Human evaluation is an expensive undertaking. It, too, has evolved over the years. Various methods to judge the content of the summaries as compared to those generated by humans have been used, beginning with SEE [12, 45] and moving to a pyramid score [43], which is the main content evaluation tool at this point. Linguistic quality was first measured by using a set of questions to rate each summary. These questions were modified, extended, and shrunk over the years and are now subsumed in two scores, one which also includes content, called "overall responsiveness", and one for linguistic quality only.

5.3.1 Human/System Summary Comparison

We present here an example of a single data set from the TAC 2011 data. The set contains 10 documents; summaries were generated by 4 humans and 43 systems. There was one baseline summarization algorithm employed. This data set was chosen specifically because it strongly demonstrates the problems of correlation between ROUGE and human evaluation. Certainly, not all data sets exhibit such wide discrepancies, but in general there is a consistent lack of correlation between the two.

Table 5.1 shows the pyramid score, overall responsiveness score, linguistic score, and ROUGE-2 score for the human summary with the highest pyramid score and the five system summaries with the highest ROUGE scores. Note that the human's ROUGE score ranges from just 40% to almost half of the system ROUGE scores, while the system pyramid scores range from barely 14% to just 26% of the human pyramid score. Clearly, in this case, ROUGE-2, which has been shown in the past to have the highest correlation with human evaluation, is lacking.

Table 5.1: Contrast between human and system scores for data set D1113C.

Summary ID	**Pyramid Score**	**Overall Responsiveness**	**Linguistic Quality**	**ROUGE-2**
human C	0.905	5	5	0.0556
system 3	0.123	3	2	0.1370
system 26	0.238	3	3	0.1369
system 35	0.153	2	2	0.1191
system 13	0.182	3	2	0.1135
system 10	0.209	2	3	0.1059

Figure 5.5 shows the human generated summary and the summary for system 26, which had the best pyramid score of the five systems with the highest ROUGE-2 scores. Note that the overall responsiveness and linguistic scores are 5, the highest possible, for the human summary while system 26 has a score of 3, the best of the five systems but clearly below the human scores. Both overall responsiveness and linguistic quality include cohesiveness, coherence, readability, and redundancy, along with more obvious grammatical issues such as run-on sentences, sentence fragments, bad punctuation and/or capitalization, lack of noun or pronoun referents, *etc.* While the system summary reads grammatically correctly, it suffers from poor ordering of the sentences that impacts the summary flow, a lack of specification for the name "Wasser", and the final sentence fragment.

A. Human Generated Summary for Data Set D1113C
Four years ago, a ship seized in Singapore carried over 6 tons of elephant ivory, the largest shipment of the contraband material since its international trade was banned in 1989. DNA identified the source of the ivory as Zambia. Since then, ivory smuggling has increased. Police seized the largest amount of ivory ever in Japan. CITES, the UN agency regulating this trade, urged African countries to overcome their divide over the ban. Positions range from requests to downgrade controls to maintaining a complete ban for 20 years. Proper elephant management can keep elephant populations stable and enable some ivory trading.

B. System 26 Summary for Data Set D1113C
Trading ivory was banned under the 1989 Convention of International Trade in Endangered Species to curb the slaughter of elephants in Africa. Using DNA analysis, the group led by Wasser determined that the tusks came from African savannah elephants similar to those found in and around the nation of Zambia. When a shipping container was seized in Singapore four years ago carrying more than SIX tons of elephant ivory inside, conservation and law enforcement agencies realized that they had intercepted the largest shipment of the contraband material since its international trade was banned in 1989. Kenya and Mali are proposing

Figure 5.5: Human and system summaries for data set D1113C.

5.4 Features

The models we employ to measure the quality of a text summary combine several features computed from the text of a summary. In this section we define these features and explain why they are included in the models. In general, our features were selected to correlate well with summary content and linguistic quality, since these are what are measured for the TAC summarization evaluation.

5.4.1 Linguistic Features

Linguistic features correspond to the grammaticality, readability, and flow, including any impact of redundancy, of the summary. These features include the number of sentences, three measures of redundancy, two measures of flow, and a measure of uniformity of sentence length. The current state of the art in automatic summarization is to extract selected sentences from the relevant document(s) to create a summary. Therefore, identifying sentences in the document is a critical early step.

Researchers at the Center for Computing Sciences have developed a sentence splitter called FASST-E (very Fast, very Accurate Sentence Splitter for Text–English) that relies on neither part-of-speech tagging nor parsing, enabling a processing speed of 1000+ sentences per second, yet is still able to execute with an error rate of less than 0.1%. FASST-E is responsible for two major tasks that contribute to the final linguistic quality of the summary. The first is, of course, the sentence splitting itself. The effort here is to ensure that all splits are made where they belong (not as simple a task as it may seem) in order to avoid both run-on sentences and sentence fragments. Run-on sentences generally cause sentences that would never be selected for a summary to be included, thereby weakening the content of the summary. Sentence fragments impact the readability, grammaticality, and continuity of the summary. The second task is to eliminate boilerplate, such as datelines, which can occur in sentences. A selected sentence which contains boilerplate has an even more negative impact on the readability, grammaticality, and continuity of the summary than sentence fragments.

The formal mathematical descriptions of our linguistic features are:

1. **Number of Sentences**: We use $-\log_2(\text{number of sentences})$. Since TAC summaries are constrained by a word limit, it is unlikely there will be too *many* sentences, but we have seen summaries comprised of just one long sentence.

2. **Redundancy Score 1**: Let $\sigma_1, \ldots, \sigma_n$ denote the singular values of the term-overlap matrix X, where $\sigma_i \geq \sigma_{i+1}$. The term-overlap matrix X is simply $(A > 0)' * (A > 0)$, where A is the term-sentence matrix and $A > 0$ denotes a logical matrix of zeros and ones. The (i, j)-entries in the term-overlap matrix are the number of terms in common in sentence i and sentence j. Redundancy score 1 is then defined as $\sum_{i=2}^{n} {\sigma_i}^2$, or the sum of the squares of all but the first singular value.

3. **Redundancy Score 2**: This is similar to the previous score; this score is calcu-

lated as $\sum_{i=3}^{n} {\sigma_i}^2$. These two redundancy scores were included to penalize summaries whose sentences overlapped too much. In the extreme case where all sentences contain the same words (in possibly different arrangements), the score would be 0.

4. **Term Entropy (Redundancy Score 3)**: Term entropy is the sample entropy of the vector of counts of term occurrences. This is calculated from the original term-sentence matrix (with zero columns removed) by dividing the column sums by the sum of all the matrix entries. Call this vector p. Then the term entropy is $-\sum_i p_i \log_2 p_i$.

5. **Sentence Entropy (Sentence Length Uniformity)**: Sentence entropy is calculated the same way as term entropy, using row sums instead of column sums. It is the sample entropy of the vector of sentence lengths. The sentence length uniformity is designed to penalize systems whose sentence lengths vary greatly. A simple example of non-uniformity of sentence length where the quality of a summary is affected was studied in [5], where it was shown that systems that use truncated sentences to end a summary have significantly lower scores than those that do not. In addition, Conroy *et al.* [8] demonstrated that this feature was more generally a useful predictor of a summary's quality.

6. **Term Overlap (Flow Score 1)**: Our first term overlap feature is computed from the term-overlap matrix, defined above in the description of Redundancy Score 1. We define the term-overlap score as the sum of the super-diagonal of this matrix, or the sum of the $(i, i+1)$-entries. The score is then the logarithm of the sum of the number of terms overlapping in each pair of adjacent sentences plus 1. We have observed that some term overlap between adjacent sentences improves readability.

7. **Normalized Term Overlap (Flow Score 2)**: The second term overlap feature is also the sum of the entries along the super-diagonal, but this time the term-overlap matrix has been symmetrically normalized first, *i.e.*, each X_{ij} has been

replaced by

$$\frac{X_{ij}}{\sigma\left(\sqrt{X_{ii}}\right)\sigma\left(\sqrt{X_{jj}}\right)}$$

where $\sigma(0) = 1$ and $\sigma(x) = x$ for $x \neq 0$.

5.4.2 Content Features

Based on the outcome of AESOP 2010, it seemed that word bigrams produced the best results in predicting the content measure of a summary. In particular, ROUGE-2 was most highly correlated with the pyramid score. As such, we focused on variations of bigram scores for content measure. In all, we investigated six variations of bigrams, the first two of which were ROUGE.

1. ROUGE-2, (R2) the consecutive bigram score.

   Let $R_n(X)$ be the ROUGE score for matching n-grams of a summary X with h human summaries denoted $M^{(j)}, j = 1, ..., h$. Then

   $$R_n(X) = \max_j \frac{\sum_{i \in N_n} \min(X_n(i), M_n^{(j)}(i))}{\sum_{i \in N_n} M_n^{(j)}(i)}$$

   where N_n is the set of $n-$grams present in the summary being scored, $X_n(i)$ is the frequency of the n-gram i in the summary and $M_n^{(j)}(i)$ is its frequency in the j-th human-generated summary.

2. ROUGE-SU4, (SU4) the bigram score that allows for a skip distance of up to 4 words.

3. Bigram coverage score (Coverage). This score is similar to ROUGE-2 but does not account for the frequency that the bigram occurs in either the human summaries or in the summary to be scored. A credit of i/n for a bigram is given if i out of n human summaries contain that bigram.

4. Unnormalized ROUGE-2 (Bigram). The score is essentially ROUGE-2 without the normalization for the length of the summaries.

5. Bigram coverage, as measured by a point to point comparison (Coverage P2P). This score is similar to the third score. However, it is computed by comparing the candidate summary to each human summary individually, as opposed to comparing it with the collection of human summaries.

6. Unnormalized ROUGE-2 as measured by a point to point comparison (Bigram P2P). This score is a point to point version of score 4.

5.5 Feature Selection and Regression Techniques

For the TAC AESOP task, we submitted four sets of predictions to each of the four different subtasks, for a total of sixteen. Each submission was based on a supervised learning algorithm performed on its own subset of features. We limited our focus here to two variations of linear regression (non-negative least squares and robust regression), and also canonical correlation, an eigenvalue method, but would like to explore other methods in the future. In particular, we chose to start with the current methods due to their ease of of application and interpretability.

In order to predict the quality of an individual summary, we took advantage of previous years' data in an interesting way. We had 13 predictors, arising from the seven linguistic features and six ROUGE-like content features. For each submission, our goal was to create a model using some subset of these features. We tested every possible combination of these 13 features. For each of the $2^{13} - 1$ combinations, we fit three models to the data. Since our ultimate goal was to predict the quality of the 2011 summaries, we trained each of these three models on the 2009 TAC data and predicted the quality of summaries in the 2010 data. For each of the three different regression methods, we used the combination of features that was best able to predict the 2010 summary quality as our feature set for predicting 2011. This gave us three sets of features, each one tailored to a particular regression method. We then used each of the three combinations of features, together with its method, to train a model on the TAC

2010 summaries. This gave us the coefficients to use for predicting the quality of the TAC 2011 summaries. The values of the coefficients are given in Tables 5.2 and 5.3 for each of our sixteen submissions.

For each subset of the covariates, we used three different methods of regression (canonical correlation, robust least squares, and non-negative least squares) to fit a predictive model. These methods are described in Section 3.5. Once we determine which subset of features has the highest possible correlation with a linear combination of the human metrics, we select that subset for evaluation on the test data.

Table 5.2: Features used when predicting scores of both human and machine summaries.

	Pyramid(8)		Responsiveness (25)		Readability(23)		Responsiveness(6)	
	canon	canon	canon	canon	robust	robust	nonneg	nonneg
Feature	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**
R2	4.8e + 1	2.4e + 1, 5.3e + 1	4.8e + 1	4.7e + 1, 8.5e + 1	4.5e + 1	6.5e + 0, -7.1e + 0		
SU4	3.7e + 1	3.9e + 1, 5.4e + 1	3.7e + 1		4.0e + 1	-6.9e + 0, 2.0e + 1	2.3e + 1	
Coverage	3.6e - 1	-3.9e - 1, -1.6e + 0	3.6e - 1	-2.4e - 1, -1.3e + 0	2.7e - 1	4.3e - 1, -2.0e - 1	1.2e - 1	7.4e - 2, 2.3e - 2
Bigrams	-4.0e - 1		-4.0e - 1		-3.6e - 1			5.9e - 3, 3.5e - 3
Bigrams P2P	5.9e - 1		5.9e - 1		3.0e + 0			
Coverage P2P	-7.2e - 1		-7.2e - 1		-3.2e + 0			
log2(1+Term Overlap)	8.9e - 2	1.9e - 1	8.9e - 2	9.3e - 2		2.9e - 2		0.0e + 0
Norm Term Overlap	-6.6e - 1	-1.0e + 0	-6.6e - 1		7.6e - 2			
Redundant 1		-2.2e - 4				-3.0e - 5		
Redundant 2				-2.0e - 4	6.7e - 4			
Term Entropy		1.0e - 1		-5.3e - 2	-5.3e - 2	2.2e - 1		
-log2(sent length)	1.2e + 0		1.2e + 0		-2.2e - 3	4.4e - 1		
Neg Sent Entropy	-9.6e - 1		-9.6e - 1		1.6e - 1	-3.8e -1		

Table 5.3: Features used when only predicting scores of machine summaries.

	Pyramid(8)		**Responsiveness (25)**		**Readability(23)**		**Responsiveness(6)**	
	robust	**robust**	**robust**	**canon**	**robust**	**robust**	**nonneg**	**nonneg**
Feature	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**
R2	1.2e + 1	2.1e + 1, 2.0e + 0		-8.9e + 1, -1.3e + 2				2.9e + 0, 0.0e + 0
SU4		3.4e + 1, 4.2e + 1	2.2e + 1	9.0e + 1, 1.0e + 2	2.2e + 1		2.0e + 1	
Coverage								
Bigrams				3.0e - 1, 4.0e - 1				9.6e - 3, 5.1e - 3
Bigrams P2P	-5.7e - 1			8.7e + 0, -1.4e + 0		2.8e + 0, -5.1e - 1		
Coverage P2P	7.3e - 1	-2.9e - 1, -5.1e - 1		-8.6e + 0, 6.8e - 1		-2.4e + 0, 4.8e - 1		
log2(1+Term Overlap)	2.7e - 1		9.3e - 2		1.4e - 1		1.0e - 1	
Norm Term Overlap	-1.2e + 0				-7.4e - 1			
Redundant 1					4.6e - 5	-1.5e - 4		
Redundant 2			5.4e - 4			3.2e - 4	6.0e - 4	
Term Entropy	-1.5e - 2		-1.0e - 1		-1.9e - 1	8.7e - 2		
-log2(sent length)				1.8e + 0	5.2e - 1	6.7e - 2		1.1e - 2
Neg Sent Entropy				-2.0e + 0	-4.6e - 1	8.4e - 2		

Each of the columns in Tables 5.2 and 5.3 describe one of our sixteen submitted sets of predictions. The numbers in the table are the coefficients used in the model named at the top of the column. In the B columns, the first six rows have two values. The first value is the coefficient for that feature in the A set and the second value is for the B set.

5.6 Results

To understand the results from the Text Analysis Conference, one more task subdivision needs to be mentioned. For both the update summaries and the initial summaries, participants submit up to four sets of score predictions (for a total of eight). However, the number is actually sixteen, since participants train each of the eight models twice. The first iteration of each is when participants are trying to predict only the scores of machine-generated summaries (called the "no models" subtask). In the second, participants train one model to simultaneously predict the scores of both human summaries and machine summaries (called the "all peers" subtask). The reason for these two tasks to exist simultaneously is that automatic evaluation systems tend to perform much better when predicting the scores of automatically generated summaries. They have much more trouble giving accurate scores to summaries written by humans, so it is interesting to see how the results change when the human summaries are included/excluded.

Figure 5.1 demonstrates why this second iteration exists. When predicting only the machine summaries, the best-fit line for ROUGE-2 and ROUGE-SU4 goes through most of the data, but does not correctly fit the human summaries. This situation improves significantly for several of our systems and is shown in Figure 5.6 with System 25 (our canonical correlation model aimed at predicting responsiveness). System 25 shrinks the gap for initial summaries and almost eliminates it for update summaries.

Figures 5.2, 5.3 and 5.6 give the Pearson correlation against overall responsiveness, pyramid scoring, and readability for the 2011 update summary tasks for the top 14 performing AESOP submissions (out of 25 total evaluation metrics). Our metrics are usually stronger for the update tasks, but in general, all of our submissions were among the top performers. We note that as indicated by the error bars, many of the best performing systems, including a number of our metrics, significantly outperformed the baselines of the three ROUGE methods in the "all peers" task. On the other hand, while ROUGE is not the best for the no models data, it is always within the 95% confidence interval of the best metric. (This confidence interval was computed using Matlab's `bootci` function, which samples with replacement from the rows of a two-column matrix consisting of the two vectors of scores. It then computes the 95% confidence interval by deriving the 2.5% and 97.5% quantile values from the distribution of correlation scores.)

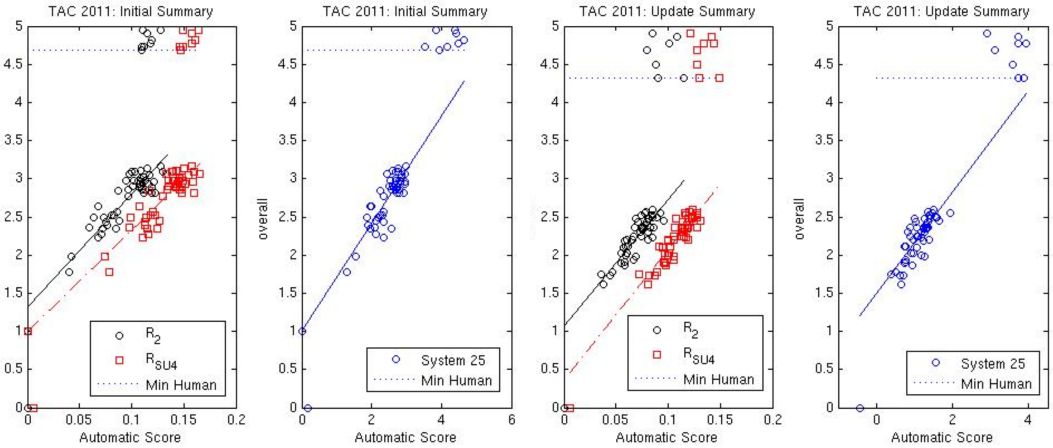


Figure 5.6: The performance of our System 25 at TAC 2011.

The two left-hand plots are ROUGE-2 and ROUGE-SU4, two of the standard baseline evaluation systems. The right-hand plots show our System 25, a canonical correlation model aimed at predicting overall responsiveness. The regression line in each plot only goes as far as the lowest-scoring human summary. System 25 closes the gap quite well for update summaries and is also a significant improvement for initial summaries.

For the update summaries (Figures 5.6, 5.2 and 5.3), there was no clear winner among our systems. Each of our four submissions was at least once our best system. System 6 (non-negative least squares model aimed at predicting responsiveness) seems to be our best for predicting the quality of machine summaries alone, and was the overall best in predicting responsiveness and pyramid for that subset. However, system 6 was twice our worst for evaluating the humans and machines together. Oliveira's system 4 (VERT-F) [11] was the best at predicting readability in the tasks with and without human summaries, and Giannakopoulos' system 18 [18] did the best at predicting responsiveness when the humans were included. Perhaps most impressively, all four of our systems (6, 8, 23 and 25) were among the top six performers (out of 25) for predicting any of the

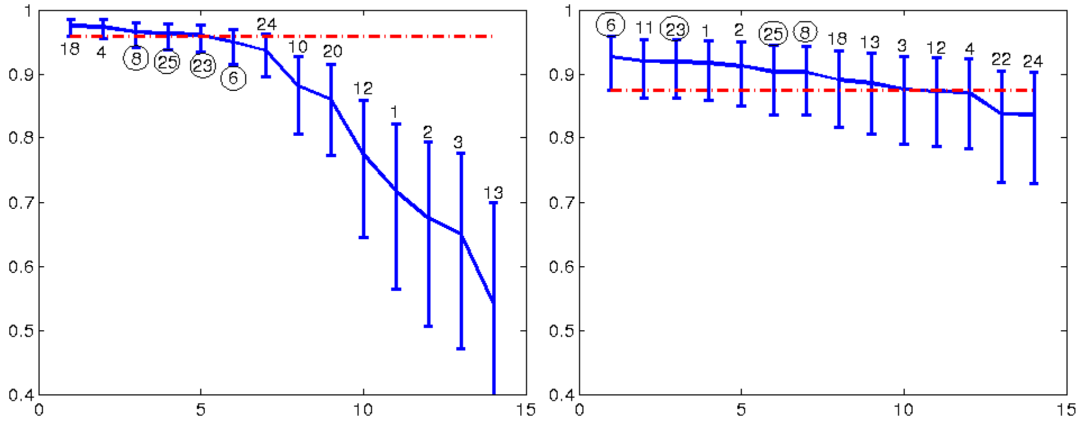


Figure 5.7: Pearson correlation with responsiveness top 14 AESOP metrics: all peers and no models (2011).
Systems 1–3 are the baseline evaluation systems (see Section 5.6). Our systems are numbers 6, 8, 23 and 25. Systems 12 (AutoSummENG) and 18 (MeMoG) are due to Giannakopoulos *et al.* [18]. System 4 is VERT-F, due to Oliveira *et al* [11]. The vertical bar shown for each system is a 95% confidence interval for its Pearson correlation with the human metric listed at the top of the figure.

three metrics with human summaries included (*i.e.*, the "all peers" task).

In order to further validate our results, we applied our sixteen tuned models to the TAC 2008 summarization data. Of our TAC 2008–2011 data, 2008 is the only other year (besides 2011) that our models did not use for training. The limitation of the 2008 data is that AESOP was not run that year, so the only systems we can compare to are the ROUGE variations. Figures 5.4, 5.8 and 5.9 show the results of these comparisons. As with 2011, our methods always outperform ROUGE in the "all peers" task, and in several cases, the differences are statistically significant. The results from the "no models" task are also similar to 2011, with our methods generally performing better than ROUGE, but not significantly so.

Baselines: Systems 1–3 are the baseline metrics and are all variants of ROUGE. In particular:

1. Baseline 1: ROUGE-2, with stemming and keeping stop words.

2. Baseline 2: ROUGE-SU4, with stemming and keeping stop words.

3. Baseline 3: Basic Elements (BE). Summaries were parsed with Minipar, and BEs were extracted and matched using the Head-Modifier criterion.

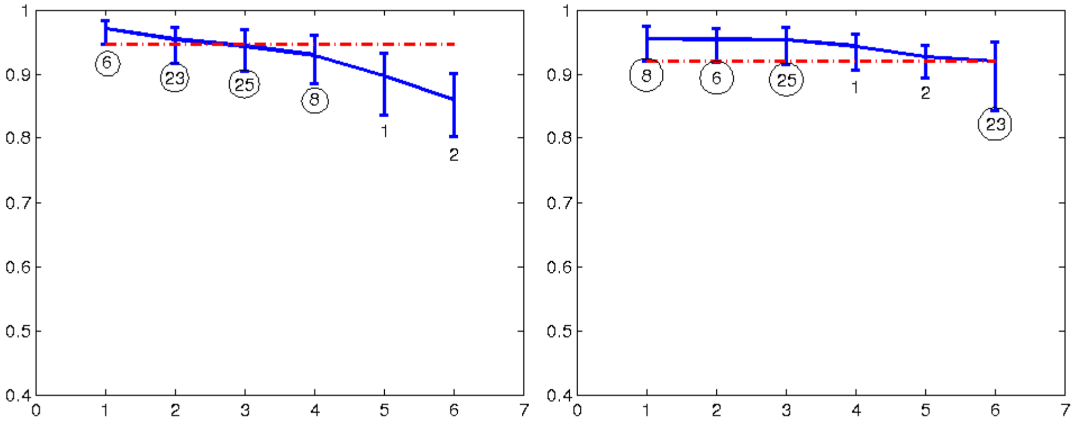


Figure 5.8: Pearson correlation with pyramid all peers and no models (2008). Systems 1 and 2 are the baseline evaluation systems (ROUGE-2 and ROUGE-SU4, respectively). Our systems are numbers 6, 8, 23 and 25. The vertical bar shown for each system is a 95% confidence interval for its Pearson correlation with the human metric listed at the top of the figure.

5.7 Conclusions and Future Work

We have demonstrated a family of metrics for estimating the quality of a text summary. The metrics are built from features which were chosen based on their correlation with human metrics. Our metrics have consistently performed very well at the Text Analysis Conference, with all four ending up in the top six or seven systems for several tasks. In particular, many of our metrics did significantly better than ROUGE in each of the "all peers" tasks (in 2008 and 2011).

At the recent 2011 INEX workshop, Tavernier and Bellot [69] reported the use of linguistic measures (such as average length of sentences, average number of syllables per word, and the frequency of monosyllabic words in the document) for the Tweet summarization task and found significant correlation with human judgment. In the future, such measures should be investigated as possible features for AESOP evaluation

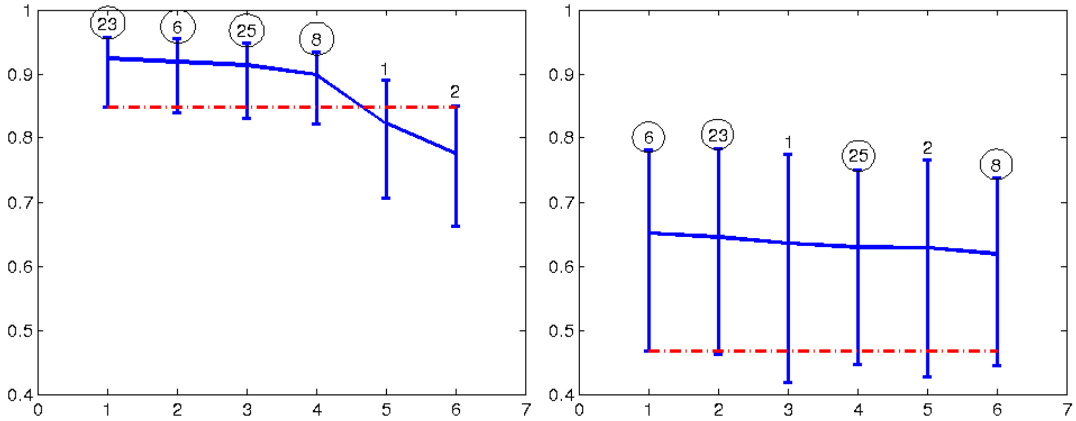


Figure 5.9: Pearson correlation with readability all peers and no models (2008). Systems 1 and 2 are the baseline evaluation systems (ROUGE-2 and ROUGE-SU4, respectively). Our systems are numbers 6, 8, 23 and 25. The vertical bar shown for each system is a 95% confidence interval for its Pearson correlation with the human metric listed at the top of the figure.

of linguistic quality. In addition, there are many other features we would like to consider for future use in AESOP, including the number of anaphora and other referential objects in a sentence. Also, many of our current features could potentially be improved with normalization and other minor adjustments.

Another avenue for future work is to optimize our models for Spearman or Kendall correlation. One way we have considered doing this is by employing a learning-to-rank method, possibly as a support vector machine.

Chapter 6

Assessing the Effect of Inconsistent Assessors on Summarization Evaluation[1]

We investigate the consistency of human assessors involved in summarization evaluation to understand its effect on system ranking and automatic evaluation techniques. Using Text Analysis Conference data, we measure annotator consistency based on human scoring of summaries for Responsiveness, Readability, and Pyramid scoring. We identify inconsistencies in the data and measure to what extent these inconsistencies affect the ranking of automatic summarization systems. We examine the stability of automatic metrics (ROUGE and CLASSY) with respect to the inconsistent assessments.

6.1 Introduction

Automatic summarization of documents is a research area that unfortunately depends on human feedback. Although attempts have been made at automating the evaluation of summaries, none is so good as to remove the need for human assessors. Human

[1] This chapter's contents first appeared in [47].

judgment of summaries, however, is not perfect either. In fact, the inconsistency of human judgments has been studied extensively, and many cases have been found where individuals give different answers to the same questions. According to Kahneman [26],

> Experienced radiologists who evaluate chest X-rays as "normal" or "abnormal" contradict themselves 20% of the time when they see the same picture on separate occasions. A study of 101 independent auditors who were asked to evaluate the reliability of internal corporate audits revealed a similar degree of inconsistency. A review of 41 separate studies of the reliability of judgments made by auditors, pathologists, psychologists, organizational managers, and other professionals suggests that this level of inconsistency is typical, even when a case is reevaluated within a few minutes.

Hence, it is not too surprising that human assessors do not always assign the same score to a summary when evaluating it on different occasions. In this chapter, we investigate two ways of measuring evaluation consistency in order to see what effect it has on summarization evaluation and training of automatic evaluation metrics.

6.2 Assessor consistency

In the Text Analysis Conference (TAC) Summarization track, participants are allowed to submit more than one run (usually two), and this option is often used to test different settings or versions of the same summarization system. In cases when the system versions are not too divergent, they sometimes produce identical summaries for a given topic. Summaries are randomized within each topic before they are evaluated, so the identical copies are usually interspersed with 40-50 other summaries for the same topic and are not evaluated in a row. Given that each topic is evaluated by a single assessor, it then becomes possible to check assessor consistency, i.e., whether the assessor judged the two identical summaries in the same way. Similar studies have been done to evaluate consistency of relevance judgments in information retrieval [59, 62].

Table 6.1: Annotator consistency in assigning Readability and Responsiveness scores and in Pyramid evaluation, as represented by Krippendorff's *alpha* for interval values, on 2011 data.

ID	Read	ID	Resp	ID	Pyr
G	0.867	G	0.931	G	0.975
D	0.866	D	0.875	D	0.970
A	0.801	H	0.808	H	0.935
H	0.783	A	0.750	A	0.931
F	0.647	F	0.720	E	0.909
C	0.641	E	0.711	C	0.886
E	0.519	C	0.490	F	0.872

For each summary, assessors conduct content evaluation according to the Pyramid framework [42, 49] and assign it Responsiveness and Readability scores[2], so assessor consistency can be checked in these three areas separately. We found between 230 (in 2009) and 430 (in 2011) pairs of identical summaries for the 2008-2011 data (given on average 45 topics, 50 runs, and two summarization conditions: main and update), giving in effect anywhere from around 30 to 60 instances per assessor per year. Using Krippendorff's *alpha* [16], we calculated assessor consistency within each year, as well as total consistency over all years' data (for those assessors who worked multiple years). Table 6.1 shows rankings of assessors in 2011, based on their Readability, Responsiveness, and Pyramid judgments for identical summary pairs (around 60 pairs per assessor). Figures 6.1 and 6.2 show a 2-way contingency table of counts of the number of times each assessor marked a single summary with each pair of scores. Larger counts are indicated with stronger colors in Figure 6.1. In Figure 6.2, however, the stronger colors indicate the severity of the mis-matched scores.

Interestingly, consistency values for Readability are lower overall than those for Responsiveness and Pyramid, even for the most consistent assessors. Given that Readability and Responsiveness are evaluated in the same way, i.e. by assigning a numerical score according to detailed guidelines, this suggests that Readability as a quality of text is inherently more vague and difficult to pinpoint.

[2] `http://www.nist.gov/tac/2011/Summarization/Guided-Summ.2011.guidelines.html`

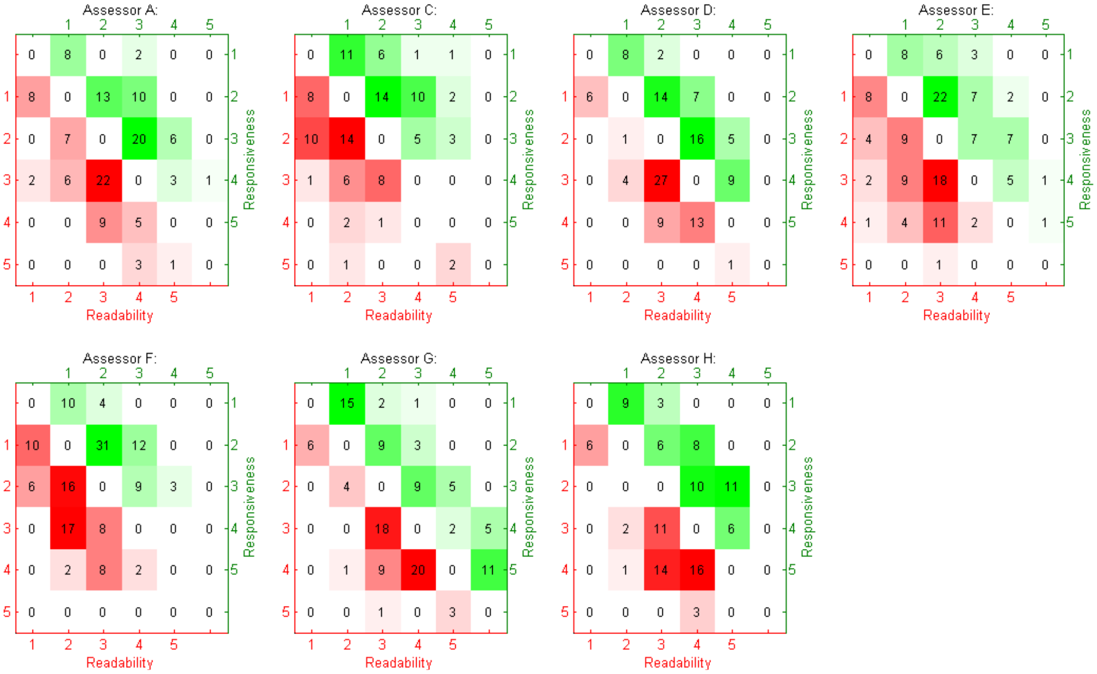
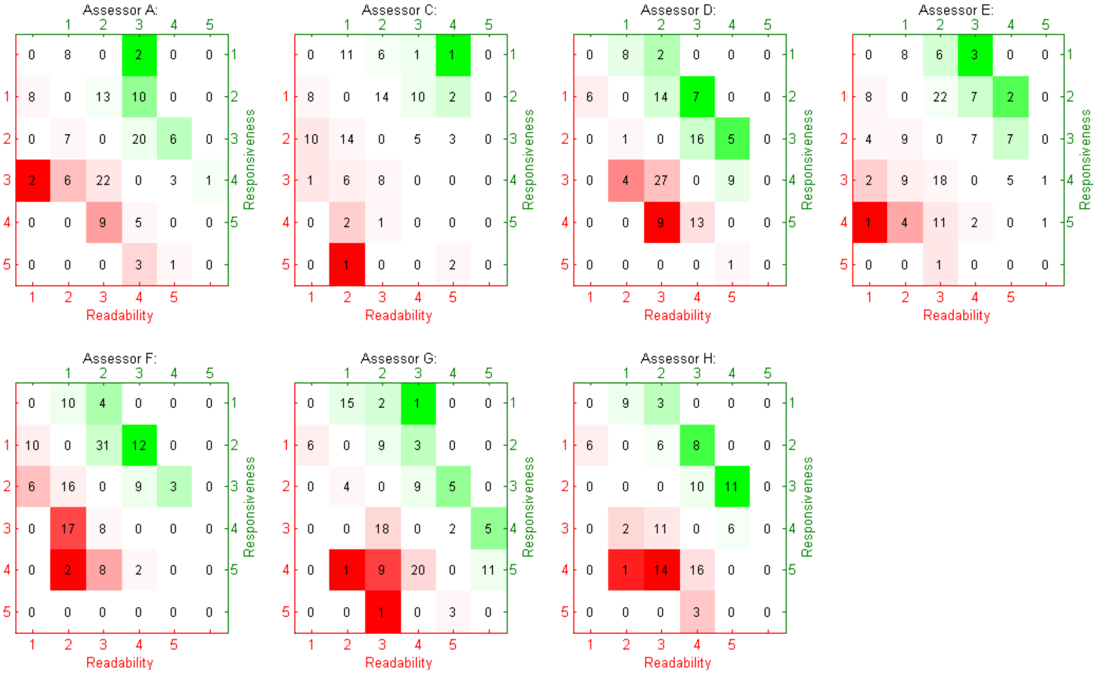


Figure 6.1: Annotator consistency in rating readability and responsiveness. Each matrix shows the number of times an assessor used each pair of values to score a single summary. The bottom left half of each matrix (shades of red/pink) shows counts for readability and the upper right (shades of green) shows counts for overall responsiveness. Intensity of color indicates larger values. A very consistent assessor has large values along the two main diagonals and smaller values elsewhere. The zeros along the actual main diagonal are just serving to separate the two halves of the matrix.

On the other hand, Pyramid consistency values are generally the highest, which can be explained by how the Pyramid evaluation is designed. Even if the assessor is inconsistent in selecting Summary Content Units (SCUs) across different summaries, as long as the total summary weight is similar, the summary's final score will be similar, too.[3] Therefore, it would be better to look at whether assessors tend to find the same SCUs (information "nuggets") in different summaries on the same topic, and whether they annotate them consistently. This can be done using the "autoannotate" function of the Pyramid process, where all SCU contributors (selected text strings) from already annotated summaries are matched against the text of a candidate (un-annotated) summary.

[3] The final score is based on total weight of all SCUs found in the summary, so the same weight can be obtained by selecting a larger number of lower-weight SCUs or a smaller number of higher-weight SCUs (or the same number of similar-weight SCUs which nevertheless denote different content).



Assessor A:

		1	2	3	4	5	
	0	8	0	2	0	0	1
1	8	0	13	10	0	0	2
2	0	7	0	20	6	0	3
3	2	6	22	0	3	1	4
4	0	0	9	5	0	0	5
5	0	0	0	3	1	0	
	1	2	3	4	5		

Assessor C:

		1	2	3	4	5	
	0	11	6	1	1	0	1
1	8	0	14	10	2	0	2
2	10	14	0	5	3	0	3
3	1	6	8	0	0	0	4
4	0	2	1	0	0	0	5
5	0	1	0	0	2	0	
	1	2	3	4	5		

Assessor D:

		1	2	3	4	5	
	0	8	2	0	0	0	1
1	6	0	14	7	0	0	2
2	0	1	0	16	5	0	3
3	0	4	27	0	9	0	4
4	0	0	9	13	0	0	5
5	0	0	0	0	1	0	
	1	2	3	4	5		

Assessor E:

		1	2	3	4	5	
	0	8	6	3	0	0	1
1	8	0	22	7	2	0	2
2	4	9	0	7	7	0	3
3	2	9	18	0	5	1	4
4	1	4	11	2	0	1	5
5	0	0	1	0	0	0	
	1	2	3	4	5		

Assessor F:

		1	2	3	4	5	
	0	10	4	0	0	0	1
1	10	0	31	12	0	0	2
2	6	16	0	9	3	0	3
3	0	17	8	0	0	0	4
4	0	2	8	2	0	0	5
5	0	0	0	0	0	0	
	1	2	3	4	5		

Assessor G:

		1	2	3	4	5	
	0	15	2	1	0	0	1
1	6	0	9	3	0	0	2
2	0	4	0	9	5	0	3
3	0	0	18	0	2	5	4
4	0	1	9	20	0	11	5
5	0	0	1	0	3	0	
	1	2	3	4	5		

Assessor H:

		1	2	3	4	5	
	0	9	3	0	0	0	1
1	6	0	6	8	0	0	2
2	0	0	0	10	11	0	3
3	0	2	11	0	6	0	4
4	0	1	14	16	0	0	5
5	0	0	0	3	0	0	
	1	2	3	4	5		

Figure 6.2: Annotator consistency in rating readability and responsiveness. Each matrix shows the number of times an assessor used each pair of values to score a single summary. The bottom left half of each matrix (shades of red/pink) shows counts for readability and the upper right (shades of green) shows counts for overall responsiveness. Intensity of color indicates very mismatched pairs of scores. A very consistent assessor has large values along the two main diagonals and smaller values elsewhere. The zeros along the actual main diagonal are just serving to separate the two halves of the matrix.

The autoannotate function works fairly well for matching between extractive summaries, which tend to repeat verbatim whole sentences from source documents. Since very few systems use paraphrasing, no attempt was made to remove such systems.

For each summary in 2008-2011 data, we autoannotated it using all remaining manually-annotated summaries from the same topic, and then we compared the resulting “autoPyramid” score with the score from the original manual annotation for that summary. Ideally, the autoPyramid score should be lower or equal to the manual Pyramid score: it would mean that in this summary, the assessor selected as relevant all the same strings as s/he found in the other summaries on the same topic, plus possibly some more information that did not appear anywhere else. If the autoPyramid score is higher than the manual Pyramid score, it means that either (1) the assessor missed

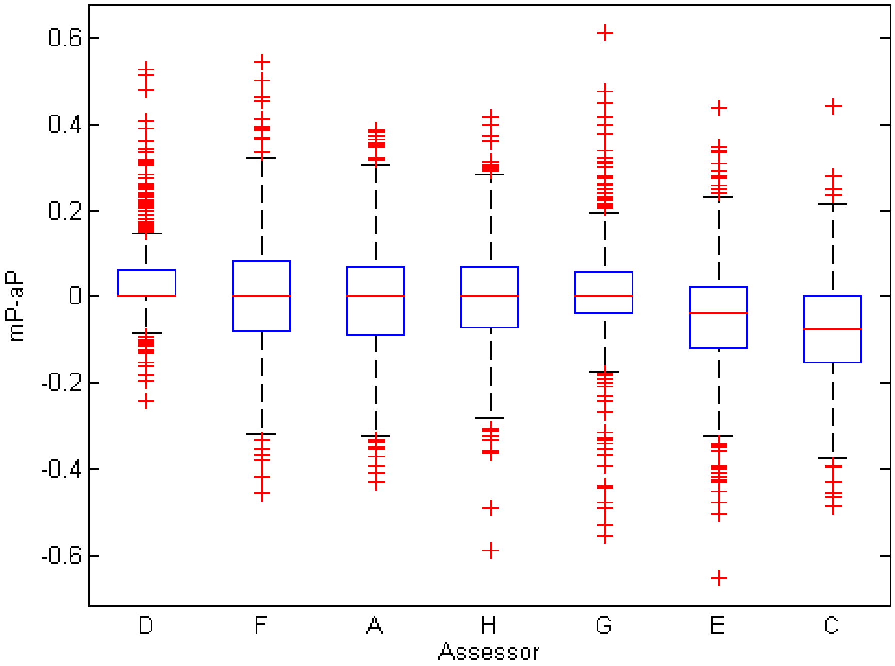


Figure 6.3: Annotator consistency in selecting SCUs in Pyramid evaluation, as represented by the difference between manual Pyramid and automatic Pyramid scores (mP-aP), on 2011 data.

relevant strings in this summary, but found them in other summaries; or (2) the strings selected as relevant elsewhere in the topic were accidental, and as such not repeated in this summary. Either way, if we then average out score differences for all summaries for a given topic, it will give us a good picture of the annotation consistency in this particular topic. Higher average autoPyramid scores suggest that the assessor was missing content, or otherwise making frequent random mistakes in assigning content. Figure 6.3 shows the macro-average difference between manual Pyramid scores and autoPyramid scores for each assessor in 2011. For the most part, it mirrors the consistency ranking from Table 6.1, confirming that some assessors are less consistent than others; however, certain differences appear: for instance, Assessor A is one of the most consistent in assigning Readability scores, but is not very good at selecting SCUs consistently. This can be explained by the fact that the Pyramid evaluation and assigning Readability scores are different processes and might require different skills and types of focus.

6.3 Impact on evaluation

Since human assessment is used to rank participating summarizers in the TAC Summarization track, we should examine the potential impact of inconsistent assessors on the overall evaluation. Because the final summarizer score is the average over many topics, and the topics are fairly evenly distributed among assessors for annotation, excluding noisy topics/assessors has very little impact on summarizer ranking. As an example, consider the 2011 assessor consistency data in Table 6.1 and Figure 6.3. If we exclude topics by the worst performing assessor from each of these categories, recalculate the summarizer rankings, and then check the correlation between the original and newly created rankings, we obtain results in Table 6.2.

Table 6.2: Correlation between the original summarizer ranking and the ranking after excluding topics by one or two worst assessors in each category.

	Pearson's r		Spearman's *rho*	
	-1 worst	-2 worst	-1 worst	-2 worst
Readability	0.995	0.993	0.988	0.986
Responsiveness	0.996	0.989	0.986	0.946
Pyramid	0.996	0.992	0.978	0.960
mP-aP	0.996	0.987	0.975	0.943

Although the impact on evaluating automatic *summarizers* is small, it could be argued that excluding topics with inconsistent human scoring will have an impact on the performance of automatic *evaluation metrics*, which might be unfairly penalized by their inability to emulate random human mistakes. Table 6.3 shows ROUGE-2 [34], one of the state-of-the-art automatic metrics used in TAC, and its correlations with human metrics, before and after exclusion of noisy topics from 2011 data. The results are fairly inconclusive: it seems that in most cases, removing topics does more harm than good, suggesting that the signal-to-noise ratio is still tipped in favor of signal. The only exception is Readability, where ROUGE records a slight increase in correlation; this is unsurprising, given that consistency values for Readability are the lowest of all categories, and perhaps here removing noise has more impact. In the case of Pyramid, there is a small gain when we exclude the single worst assessor, but excluding two assessors results in a decreased correlation, perhaps because we remove too much valid information at the same time.

Table 6.3: Correlation between the summarizer rankings according to ROUGE-2 and human metrics, before and after excluding topics by one or two worst assessors in that category.

	Readability	Responsiveness	Pyramid	mP-aP
before	0.705	0.930	0.954	0.954
-1 worst	0.718	0.921	0.961	0.942
-2 worst	0.718	0.904	0.952	0.923

A different picture emerges when we examine how well ROUGE-2 can predict human scores on the *summary* level. We pooled together all summaries annotated by each particular assessor and calculated the correlation between ROUGE-2 and this assessor's manual scores for individual summaries. Then we calculated the mean correlation over all assessors. Unsurprisingly, inconsistent assessors tend to correlate poorly with automatic (and therefore always consistent) metrics, so excluding one or two worst assessors from each category increases ROUGE's average per-assessor summary-level correlation, as can be seen in Table 6.4. The only exception here is when we exclude assessors based on their autoPyramid performance: again, because inconsistent SCU selection does not necessarily translate into inconsistent final Pyramid scores, excluding those assessors does not do much for ROUGE-2.

Table 6.4: Correlation between ROUGE-2 and human metrics on a summary level before and after excluding topics by one or two worst assessors in that category.

	Readability	Responsiveness	Pyramid	mP-aP
before	0.579	0.694	0.771	0.771
-1 worst	0.626	0.695	0.828	0.752
-2 worst	0.628	0.721	0.817	0.741

6.4 Impact on training

Another area where excluding noisy topics might be useful is in training new automatic evaluation metrics. To examine this issue we turned to CLASSY [8, 57], an automatic evaluation metric submitted to TAC each year from 2009-2011. CLASSY consists of four different versions, each aimed at predicting a particular human evaluation score. Each version of CLASSY is based on one of three regression methods: robust regression, non-negative least squares, or canonical correlation. The regressions are calculated based on a collection of linguistic and content features, derived from the summary to be scored.

CLASSY requires two years of marked data to score summaries in a new year. In order to predict the human metrics in 2011, for example, CLASSY uses the human ratings from 2009 and 2010. It first considers each subset of the features in turn, and using each of the regression methods, fits a model to the 2009 data. The subset/method combination that best predicts the 2010 scores is then used to predict scores for 2011. However, the model is first re-trained on the 2010 data to calculate the coefficients to be used in predicting 2011.

First, we trained all four CLASSY versions on all available 2009-2010 topics, and then trained again excluding topics by the most inconsistent assessor(s). A different subset of topics was excluded depending on whether this particular version of CLASSY was aiming to predict Responsiveness, Readability, or the Pyramid score. Then we tested CLASSY's performance on 2011 data, ranking either automatic summarizers (NoModels case) or human and automatic summarizers together (AllPeers case), separately for main and update summaries, and calculated its correlation with the metrics it was aiming to predict. Table 6.5 shows the result of this comparison. For Pyramid, (a) indicates that excluded topics were selected based on Krippendorff's *alpha*, and (b) indicates that topics were excluded based on their mean difference between manual and automatic Pyramid scores.

The results are encouraging; it seems that removing noisy topics from training data

Table 6.5: Correlations between CLASSY and human metrics on 2011 data (main and update summaries), before and after excluding most inconsistent topic from 2009-2010 training data for CLASSY.

	NoModels		AllPeers	
	main	update	main	update
Pyramid				
CLASSY1_Pyr	0.956	0.898	0.945	0.936
CLASSY1_Pyr_new (a)	0.950	0.895	0.932	**0.955**
CLASSY1_Pyr_new (b)	**0.960**	**0.900**	0.940	**0.955**
Responsiveness				
CLASSY2_Resp	0.951	0.903	0.948	0.963
CLASSY2_Resp_new	**0.954**	**0.907**	**0.973**	0.950
CLASSY4_Resp	0.951	0.927	0.830	0.949
CLASSY4_Resp_new	0.943	**0.928**	**0.887**	0.946
Readability				
CLASSY3_Read	0.768	0.705	0.844	0.907
CLASSY3_Read_new	**0.793**	**0.721**	**0.858**	0.906

does improve the correlations with manual metrics in most cases. The greatest increase takes place in CLASSY's correlations with Responsiveness for main summaries in AllPeers case, and for correlations with Readability. While none of the changes are large enough to achieve statistical significance, the pattern of improvement is fairly consistent.

6.5 Conclusions

We investigated the consistency of human assessors in the area of summarization evaluation. We considered two ways of measuring assessor consistency, depending on the metric, and studied the impact of consistent scoring on ranking summarization systems and on the performance of automatic evaluation systems. We found that summarization system ranking, based on scores for multiple topics, was surprisingly stable and did not change significantly when several topics were removed from consideration. However, on a summary level, removing topics scored by the most inconsistent assessors helped ROUGE-2 increase its correlation with human metrics. In the area of training automatic metrics, we found some encouraging results; removing noise from the training data allowed most CLASSY versions to improve their correlations with the manual metrics that they were aiming to model.

Chapter 7

A Decade of Automatic Content Evaluation of News Summaries: Reassessing the State of the Art[1]

How good are automatic content metrics for news summary evaluation? Here we provide a detailed answer to this question, with a particular focus on assessing the ability of automatic evaluations to identify statistically significant differences present in manual evaluation of content. Using four years of data from the Text Analysis Conference, we analyze the performance of eight ROUGE variants in terms of accuracy, precision and recall in finding significantly different systems. Our experiments show that some of the neglected variants of ROUGE, based on higher order n-grams and syntactic dependencies, are most accurate across the years; the commonly used ROUGE-1 scores find too many significant differences between systems which manual evaluation would deem comparable. We also test combinations of ROUGE variants and find that they considerably improve the accuracy of automatic prediction. Finally, we discuss the importance of reporting significance when claiming a new system is better than the current state of the art.

[1] This chapter's contents first appeared in [56].

7.1 Introduction

ROUGE [35] is a suite of automatic evaluations for summarization and was introduced a decade ago as a reasonable substitute for costly and slow human evaluation. The scores it produces are based on n-gram or syntactic overlap between an automatic summary and a set of human reference summaries. However, the field does not have a good grasp of which of the many evaluation scores is most accurate in replicating human judgments. This state of uncertainty has led to problems in comparing published work, as different researchers choose to publish different variants of scores.

In this paper we reassess the strengths of ROUGE variants using the data from four years of Text Analysis Conference (TAC) evaluations, 2008 to 2011. To assess the performance of the automatic evaluations, we focus on determining statistical significance[2] between systems, where the gold-standard comes from comparing the systems using manual pyramid and responsiveness evaluations. In this setting, computing correlation coefficients between manual and automatic scores is not applicable as it does not take into account the statistical significance of the differences nor does it allow the use of more powerful statistical tests which use pairwise comparisons of performance on individual document sets. Instead, we report on the accuracy of decisions on pairs of systems, as well as the precision and recall of identifying pairs of systems which exhibit statistically significant differences in content selection performance.

7.2 Background

During 2008–2011, automatic summarization systems at TAC were required to create 100-word summaries. Each year there were two multi-document summarization subtasks, the initial summary and the update summary, usually referred to as task A and task B, respectively. The test inputs in each consisted of about 10 documents and the type of summary varied between query-focused and guided. There are between 44 and

[2] For the purpose of this study, we define a difference as significant when the test statistic attains a value corresponding to a p-value less than 0.05.

48 test inputs on which systems are compared for each task.

In 2008 and 2009, task A was to produce a query-focused summary in response to a user information need stated both as a brief statement and a paragraph-long description of the information the user seeks to find. In 2010 and 2011 task A was "guided summarization", where the test inputs came from a small set of predefined domains. These domains included accidents and natural disasters, attacks, health and safety, endangered resources, investigations and trials. Systems were provided with a list of important aspects of information for each domain and were asked to cover as many of these aspects as possible. The writers of the reference summaries for evaluation were given similar instructions. In all four years, task B was to produce an update summary for each of the inputs given in task A (query-focused or guided). In each case, a new, subsequent set of documents related to the topic of the respective test set for task A was provided to the system. The task was to generate an update summary aimed at a user who has already read all documents in the inputs for task A.

The two manual evaluation approaches used in TAC 2008–2011 are modified pyramid [43] and overall responsiveness. The pyramid method requires several reference summaries for each input. These are manually analyzed to discover content units based on meaning rather than specific wording. Each content unit is assigned a weight equal to the number of reference summaries that included that content unit. The modified pyramid score is defined as the sum of weights of the content units in the summary normalized by the weight of an ideally informative summary which expresses n content units, where n is equal to the average of content units in the reference summaries. Responsiveness, on the other hand, is based on direct human judgments, without the need for reference summaries. Assessors are presented with a statement of the user's information need and the summary they need to evaluate. Then they rate how well they think the summary responds to the information need contained in the topic statement. Responsiveness was rated on a ten-point scale in 2009, and on a five-point scale in all other years.

Table 7.1: Number of pairs of significantly different systems among the top 30 across the years. There is a total of 435 pairs ($435 = \binom{30}{2}$) in each year.

Year	Pyr A	Pyr B	Resp A	Resp B
2008	82	109	68	105
2009	146	190	106	92
2010	165	139	150	128
2011	39	83	5	11

For each sub-task during 2008–2011, we analyze the performance of only the top 30 systems, which roughly corresponds to the systems that performed better than or around the median according to each manual metric. Table 7.1 gives the number of significant differences among the top 30 participating systems. These significances were calculated using a Wilcoxon signed-rank test on two vectors of human-produced scores. In the case of responsiveness, ties were removed from the samples and the test was conducted on the remaining observations. These tests used the cutoffs provided by large sample normal theory. We keep only the best performing systems for the analysis because we are interested in studying how well automatic evaluation metrics can correctly compare very good systems.

7.3 Which ROUGE is best?

In this section, we study the performance of several ROUGE variants, including ROUGE-n, for $n = 1, 2, 3, 4$, ROUGE-L, ROUGE-W-1.2, ROUGE-SU4, and ROUGE-BE-HM [24]. ROUGE-n measures the n-gram recall of the evaluated summary compared to the available reference summaries. ROUGE-L is the ratio of the number of words in the longest common subsequence between the reference and the evaluated summary and the number of words in the reference. ROUGE-W-1.2 is a weighted version of ROUGE-L. ROUGE-SU4 is a combination of skip bigrams and unigrams, where the skip bigrams are formed for all words that appear in the text with no more than four intervening words in between. ROUGE-BE-HM computes recall of dependency syntactic relations between the summary and the reference.

To evaluate how well an automatic evaluation metric reproduces human judgments, we use prediction *accuracy* similar to [46]. For each pair of systems in each subtask, we compare the results of two Wilcoxon signed-rank tests, one using the manual evaluation scores for each system and one using the automatic evaluation scores for each system. We use the Wilcoxon test as it was demonstrated in Chapter 4 to give slightly more power (not necessarily significant) against alternatives of interest than unpaired tests. As reported by [77], other tests such as randomized testing, may also be appropriate. There is considerable variation in system performance for different inputs [41] and paired tests remove the effect of the input. Both Wilcoxon tests removed ties and operated on the reduced sample size. Significance cutoffs came from the large sample normal distribution theory, conditioned on the reduced sample size after ties were removed. The accuracy then is simply the percent agreement between the results of these two tests.

As can be seen in Table 7.1, the manual evaluation metrics often did not show many significant differences between systems.[3] Thus, it is clear that the percent agreement will be high for an approach for automatic evaluation that always predicts zero significant differences. As traditionally done when dealing which such skewed distributions of classes, we also examine the *precision* and *recall* with respect to finding significant differences of several ROUGE variants, to better assess the quality of their prediction. To identify a measure that is strong at predicting both significant and non-significant differences we compute balanced accuracy, the mean of the accuracy of predicting significant differences and the accuracy of predicting no significant difference. More generally, one could define a utility function which gives costs associated with errors and benefits to correct prediction. Balanced accuracy weighs all errors as equally bad and all correct

[3] This is a somewhat surprising finding which may warrant further investigation. One possible explanation is that different systems generate similar summaries. Recent work has shown that this is unlikely to be the case because the collection of summaries from several systems indicates better what content is important than the single best summary [37]. The short summary length for which the summarizers are compared may also contribute to the fact that there are few significant difference. In early NIST evaluations manual evaluations could not distinguish automatic and human summaries based on summaries of length 50 and 100 words and there were more significant differences between systems for 200-word summaries than for 100-word summaries [40].

Table 7.2: Sample contingency table for evaluating binary predictions.

	Predicted Sig.	Predicted Non-Sig.
Actual Sig.	TP	FN
Actual Non-Sig.	FP	TN

prediction as equally good [73].

To further understand the four measures we use throughout the rest of this chapter, consider the sample contingency table in Table 7.2. The values in the body of Table 7.2 are true positives (TP), false negatives (FN), false positives (FP), and true negatives (TN). The four measures are then calculated as:

1. Accuracy $= \frac{TP+TN}{TP+FP+FN+TN}$

2. Precision $= \frac{TP}{TP+FP}$

3. Recall $= \frac{TP}{TP+FN}$

4. Balanced Accuracy $= \frac{1}{2}(\frac{TP}{TP+FN} + \frac{TN}{TN+FP})$

Each of these four measures for judging the performance of ROUGE variants has direct intuitive interpretation, unlike other opaque measures such as correlation coefficients and F-measure which have formal definitions which do not readily yield to intuitive understanding.

Few prior studies have taken statistical significance into account during the assessment of automatic metrics for evaluation. For this reason we first briefly discuss ROUGE accuracy without taking significance into account. In this special case, agreement simply means that the automatic and manual evaluations agree on which of two systems is better, based on each system's average score for all test inputs for a given task. It is very rare that the average scores of two systems are equal, so there is always a better system in each pair, and random prediction would have 50% accuracy.

Many papers do not report the significance of differences in ROUGE scores (for the ROUGE variant of their choice), but simply claim that their system X with higher

Table 7.3: Accuracy, Precision, Recall, and Balanced Accuracy of each ROUGE variant, averaged across all eight tasks in 2008-2011, with and (without) significance.

	Responsiveness				Pyramid			
Metric	Acc	P	R	BA	Acc	P	R	BA
R1	0.58 (0.61)	0.24	0.64	0.57	0.62 (0.66)	0.37	0.67	0.61
R2	0.64 (0.63)	0.28	0.60	0.59	0.68 (0.69)	0.43	0.63	0.64
R3	0.70 (0.63)	0.31	0.48	0.60	0.73 (0.68)	0.49	0.53	0.66
R4	0.73 (0.64)	0.33	0.40	0.60	0.74 (0.65)	0.50	0.45	0.65
RL	0.50 (0.59)	0.20	0.56	0.54	0.54 (0.63)	0.29	0.60	0.55
R-SU4	0.61(0.62)	0.26	0.61	0.58	0.65 (0.68)	0.40	0.65	0.63
R-W-1.2	0.52(0.62)	0.21	0.54	0.55	0.57(0.64)	0.32	0.62	0.57
R-BE-HM	0.70 (0.63)	0.30	0.49	0.59	0.74(0.68)	0.49	0.56	0.66

average ROUGE score than system Y is better than system Y. Table 7.3 lists the average accuracy with significance taken into account and then in parentheses, accuracy without taking significance into account. The data demonstrate that the best accuracy of the eight ROUGE metrics is a meager 64% for responsiveness when significance is not taken into account. So the conclusion about the relative merit of systems would be different from that based on manual evaluation in one out of three comparisons. However, the best accuracy rises to 73% when significance is taken into account; an incorrect conclusion will be drawn in one out of four comparisons. The reduction in error is considerable.

Furthermore, ROUGE-3 and ROUGE-4, which are rarely reported, are among the most accurate. Note also, these results differ considerably from those reported by [46], where ROUGE-2 was shown to have accuracy of 81% for responsiveness and 89% for pyramid. The wide differences are due to the fact we are only considering systems which scored in the top 30. This illustrates that our automatic metrics are not as good at discriminating systems near the top. These findings give strong support for the idea of requiring authors to report the significance of the difference between their summarization system and the chosen baseline; the conclusions about relative merits of the system would be more similar to those one would draw from manual evaluation.

In addition to accuracy, Table 7.3 gives precision, recall and balanced accuracy for each of the eight ROUGE measures when significance is taken into account. ROUGE-1 is arguably the most widely used score in the literature and Table 7.3 reveals an interesting property: ROUGE-1 has high recall but low precision. This means that it reports many significant differences, most of which do not exist according to the manual evaluations.

Balanced accuracy helps us identify which ROUGE variants are most accurate in finding statistical significance and correctly predicting that two systems are not significantly different. For the pyramid evaluation, the variants with best balanced accuracy (66%) are ROUGE-3 and ROUGE-BE, with ROUGE-4 just a percent lower at 65%. For responsiveness the configuration is similar, with ROUGE-3 and ROUGE-4 tied for best (60%), and ROUGE-BE just a percent lower.

The good performance of higher-order n-grams is quite surprising because these are practically never used for reporting results in the literature. Based on our results however, they are much more likely to accurately reproduce conclusions that would have been drawn from manual evaluation of top-performing systems.

7.4 Multiple hypothesis tests to combine ROUGE variants

We now consider a method to combine multiple evaluation scores in order to obtain a stronger ensemble metric. The idea of combining ROUGE variants has been explored in the prior literature. [5], for example, proposed taking linear combinations of ROUGE metrics. This approach was extended by [57] by including measures of linguistic quality. Recently, [1] applied the "heterogeneity principle" and combined ROUGE scores to improve the *precision* relative to a human evaluation metric. Their results demonstrate that a consensus among ROUGE scores can predict more accurately if an improvement in a human evaluation metric will be achieved.

Along the lines of these investigations, we examine the performance of a simple combination of variants: Call the difference between two systems significant only when *all* the variants in the combination indicate significance. As in the section above, a paired Wilcoxon signed-rank test using normal theory large sample approximations is used to determine the level of significance.

Table 7.4: Accuracy, Precision, Recall, and Balanced Accuracy of each ROUGE combination on TAC 2008-2010 pyramid.

ROUGE Combination	Acc	Prec	Rec	BA
R1_R2_R4_RBE	0.76	0.77	0.36	0.76
R1_R4_RBE	0.76	0.76	0.36	0.76
R2_R4_RBE	0.76	0.74	0.40	0.75
R4_RBE	0.76	0.73	0.41	0.75
R1_R2_R4	0.76	0.71	0.40	0.74
R1_R4	0.75	0.70	0.40	0.73
R2_R4	0.75	0.68	0.44	0.73
R1_R2_RBE	0.75	0.66	0.48	0.72
R2_RBE	0.75	0.64	0.52	0.72
R4	0.74	0.62	0.47	0.70
R1_RBE	0.74	0.62	0.49	0.70
R1_R2	0.73	0.57	0.62	0.70
RBE	0.73	0.57	0.58	0.68
R2	0.71	0.53	0.69	0.68
R1	0.62	0.43	0.69	0.63

We considered all possible combinations of four ROUGE metrics that exhibited good properties in the analyses presented so far: ROUGE-1 (because of its high recall), ROUGE-2 (because of high accuracy when significance is not taken into account) and ROUGE-4 and ROUGE-BE, which showed good balanced accuracy.

The performance of these combinations for reproducing the decisions in TAC 2008-2010 based on the pyramid[4] evaluation are given in Table 7.4. The best balanced accuracy (76%) is for the combination of all four variants. As more variants are combined, precision increases but recalls drops.

[4] The ordering of the metric combinations relative to responsiveness was almost identical to the ordering relative to the pyramid evaluation, and precision and recall exhibited the same trend as more metrics were added to the combination.

Table 7.5: Best performing AESOP systems from TAC 2011; Scores within the 95% confidence interval of the best are in bold face.

	Pyramid A				Pyramid B				Responsiveness A				Responsiveness B			
Evaluation Metric	Acc	P	R	BA	Acc	P	R	BA	Acc	P	R	BA	Acc	P	R	BA
CLASSY1	0.60	0.02	**0.60**	0.50	**0.84**	0.03	0.18	0.50	0.61	0.14	0.64	0.54	0.70	0.21	0.22	0.52
DemokritosGR1	0.59	0.01	0.20	0.50	0.79	0.07	0.55	0.53	0.66	0.18	**0.79**	0.58	0.64	0.17	0.24	0.49
uOttawa3	0.44	0.01	**0.60**	0.50	0.48	0.02	0.36	0.50	0.52	0.13	0.77	0.55	0.43	0.13	0.36	0.46
DemokritosGR2	0.78	0.01	0.20	0.50	0.76	0.06	0.55	0.52	0.76	0.23	0.69	0.60	0.67	0.22	0.29	0.52
C-S-IIITH4	0.69	0.01	0.20	0.50	0.77	0.07	0.64	0.53	0.82	**0.29**	0.74	**0.63**	0.60	0.15	0.24	0.47
C-S-IIITH1	0.60	0.01	0.40	0.50	0.70	0.06	**0.82**	0.53	0.69	0.20	**0.79**	0.59	0.60	0.22	**0.42**	0.52
BEwT-E	0.73	0.01	0.20	0.50	0.80	0.01	0.09	0.49	0.79	0.25	0.72	**0.61**	0.72	**0.31**	0.39	**0.58**
R1-R2-R4-RBE	**0.89**	**0.40**	0.44	**0.67**	0.76	0.27	0.17	0.55	**0.88**	0.00	0.00	0.49	**0.91**	0.03	0.09	0.50
R1-R4-RBE	**0.89**	**0.40**	0.44	**0.67**	0.77	**0.35**	0.24	**0.59**	**0.88**	0.00	0.00	0.49	0.90	0.03	0.09	0.50
All ROUGEs	**0.89**	**0.40**	0.44	**0.67**	0.75	0.26	0.16	0.54	**0.88**	0.00	0.00	0.49	**0.91**	0.04	0.09	0.51

7.5 Comparison with automatic evaluations from AESOP 2011

In 2009-2011, TAC ran the task of Automatically Evaluating Summaries of Peers (AESOP), to compare automatic evaluation methods for automatic summarization. Here we show how the submitted AESOP metrics compare to the best ROUGE variants that we have established so far. We report the results on 2011 only, because even when the same team participated in more than one year, the metrics submitted were different and the 2011 results represent the best effort of these teams. However, as we saw in Table 7.1, in 2011 there were very few significant differences between the top summarization systems. In this sense the tasks that year represent a challenging dataset for testing automatic evaluations.

The results for the best AESOP systems (according to one or more measures), and the corresponding results for the ROUGE combinations are shown in Table 7.5. These AESOP systems are: CLASSY1 [7, 57], DemokritosGR1 and 2 [18, 19], uOttawa3 [29], C-S-IITH1 and 4 [31, 32], and BEwT-E [70].[5] The combination metrics achieve the highest accuracy by generally predicting correctly when there are no significant differences

[5] To perform the comparison in the table the scores for each system and document set were needed. Some systems have changed after TAC 2011, but the data needed for these comparisons were not available. BEwT-E did not participate in AESOP 2011 and these data were provided by Stephen Tratz. Special thanks to Stephen for providing these data.

between the systems. In addition, for 2008-2010, where far more differences between systems occur, the results of Table 7.4 show the combination metrics outperformed use of a single metric and are competitive with the best metrics of AESOP 2011. Thus, the combination metrics have the ability to discriminate under both conditions giving good prediction of human evaluation. than system Y when the opposite is true according to the manual evaluation.

7.6 Conclusion

We have tested the best-known automatic evaluation metrics (ROUGE) on several years of TAC data and compared their performance with recently developed AESOP metrics. We discovered that some of the rarely used variants of ROUGE perform surprisingly well, and that by combining different ROUGEs together, one can create an evaluation metric that is extremely competitive with metrics submitted to the latest AESOP task. Our results were reported in terms of several different measures, and in each case, compared how well the automatic metric predicted significant differences found in manual evaluation. We believe strongly that developers should include statistical significance when reporting differences in ROUGE scores of theirs and other systems, as this improves the accuracy and credibility of their results. Significant improvement in multiple ROUGE scores is a significantly stronger indicator that the developers have made a noteworthy improvement in text summarization. Systems that report significant improvement using a combination of ROUGE-BE (or its improved version BEwT-E) in conjunction with ROUGE-1, 2, and 4, are more likely to give rise to summaries that humans would judge as significantly better.

Chapter 8

Conclusions and Future Work

8.1 Conclusions

This dissertation investigated many aspects of summarization evaluation and aimed to bring statistical ideas and theory to the field. One over-arching goal of this study was to carefully consider the ways in which statistical testing could be applied to scores from summarization tasks. The data used in these experiments presented a unique challenge in that much of it was ordinal scale Likert scores, and these do not conform to many of the assumptions used in most statistical tests. Options for dealing with this type of data were explored and discussed, and still more investigations must be done.

The framework that encapsulates the evaluation of automatic summarization was studied in detail. The power of different statistical tests was evaluated and demonstrated. A family of automatic summarization metrics was developed and proven to perform very well in the TAC evaluation task. The consistency of human assessors was investigated in great detail in Chapter 6, and the effect inconsistent assessors can have on the evaluation of summarization was studied as well. Finally, it was suggested that researchers should additionally report significance results with their improvement numbers, and reasons why were demonstrated.

8.2 Future Work

In Chapter 7, I argued the importance of reporting the statistical significance of a favorable comparison between one's system and the current state of the art. However, measures of statistical significance have their own difficulties in interpretation. It would interesting to investigate a larger range of possible metrics to report. Perhaps the magnitude of the difference in ROUGE scores, for example, would be meaningful. A further study might be then needed on how to interpret ROUGE score improvements.

Also in Chapter 7, we investigated the usefulness of reporting how often a candidate metric agrees with the human scores when used to decide statistical significance. The challenging part about this metric is the high variability of the linguistic complexity of documents. The percentage agreement on one task in one year will probably not be comparable to the percent agreement calculated elsewhere. An interesting way to make sense of all this might be to attempt to predict how difficult a document corpus will be to summarize. In some sense, this would be looking at the same matrix of scores but from the other direction. This study would be immediately feasible due to the availability of data that has not really been used for this purpose. In fact, studies [3] have been done using document similarity, something that could be useful in this area.

I mentioned in Chapter 5 that NIST calculates several correlations to measure the performance of the automatic evaluation metrics. The metrics get graded on how well their ranking of the machine systems correlates with the ranking produced by human scores. I did not report the Kendall tau correlation [28] in this paper, but it is used at the workshop. However, due to the potential difficulties [60] with Kendall's tau, it would be interesting to test the performance and interpretability of τ_{ap} [78], a variant that puts more weight/emphasis on errors made near the top of a list. For example, if we ranked summarization systems according to their overall responsiveness scores, we could then test an automatic evaluation metric like ROUGE-2 and see if the systems are ranked the same way with the new scores. Calculating τ_{ap} would be a reasonable way to judge an evaluation metric.

It seems sensible to me that a paired test like the Wilcoxon signed rank test should preform better than a t-test (paired or unpaired) on Likert-scale data, but it does not seem to be the case. A further study to investigate why exactly this happens could be beneficial to the summarization community, especially since their data so often takes this form.

Other studies have also shown unexpected results of this sort. For example, Smucker, Allan, and Carterette [65] compared a randomization test with the paired t-test and a bootstrap test at very small sample sizes (as few as 10 topics). They showed that the three tests agree at a very high rate, but that as the number of topics decreases, the tests start to disagree.

Appendix A

Distribution of Wilcoxon Signed Rank Statistic

In this section, the asymptotic distribution of the Wilcoxon signed rank statistic is derived in terms of U-statistics in the general case where the distribution of the underlying random variables may have jumps. Similar derivations exist in the literature for the specific cases where the distributions are assumed continuous. For the case with ties, I have never found the place where they are worked out carefully and completely (although most of the ideas used in Section A.1 can be found in [14] and [15]). The purpose of this section, therefore, is to provide such derivations for the two different cases that arise: the case where non-zero ties are assumed impossible (Section A.1), and the case where non-zero ties are not impossible (Section A.2). In both instances, it is assumed there are no "zero ties" in the data, i.e., that $X_i = Y_i$ since this type of tie would be removed before this point in the analysis.[1] The notation $I(X > 0)$ is meant to be the indicator function that takes the value 1 if $X > 0$ and the value 0 otherwise.

[1] This procedure was outlined in Section 3.2. The first step of the signed-rank test is to remove (X_i, Y_i) from the X and Y vectors if $X_i = Y_i$ and reduce the sample size accordingly.

A.1 The No-Ties Case [2]

Define the Wilcoxon signed rank test statistic as:

$$W_n^+ = \sum_{i=1}^{n} R_i^+ I(Z_i > 0),$$

where R_i^+ is the rank of $|Z_i|$ among $|Z_1|, \dots, |Z_n|$. The absolute rank of Z_i can be written as $R_i^+ = 1 + \sum_{j=1}^{n} I(|Z_j| < |Z_i|)$ since the RHS counts how many $|Z_j|$ are less than $|Z_i|$, and then adds 1. Using this, we can replace R_i^+ in the formula above and write

$$W_n^+ = \sum_{i=1}^{n} \left(1 + \sum_{j=1}^{n} I(|Z_j| < |Z_i|) \right) I(Z_i > 0). \tag{A.1.0.1}$$

Using the distributive law, this becomes

$$W_n^+ = \sum_{i=1}^{n} I(Z_i > 0) + \sum_{i=1}^{n} I(Z_i > 0) \sum_{j=1}^{n} I(|Z_j| < |Z_i|),$$

which can be re-written as

$$W_n^+ = \sum_{i=1}^{n} I(Z_i > 0) + \sum_{i=1}^{n} \sum_{j=1}^{n} I(Z_i > 0) I(|Z_j| < |Z_i|).$$

But of course, since $|Z_j| < |Z_i|$ never happens if $i = j$, we can reduce the double sum to only summing over values of i and j with $i \neq j$. Also, the two indicator functions inside the sum can be combined into one. Hence,

$$W_n^+ = \sum_{i=1}^{n} I(Z_i > 0) + \sum_{i \neq j} I(Z_i > 0, |Z_j| < |Z_i|). \tag{A.1.0.2}$$

We can now break up $Z_i + Z_j > 0$ into two cases, depending on whether $|Z_i| > |Z_j|$ or $|Z_j| > |Z_i|$. If $|Z_i| > |Z_j|$, then $Z_i + Z_j > 0$ if and only if $Z_i > 0$. In the opposite case, when $|Z_i| < |Z_j|$, we have $Z_i + Z_j > 0$ if and only if $Z_j > 0$. The two cases can be summarized neatly as one indicator equalling the sum of the other two:

$$I(Z_i + Z_j > 0) = I(|Z_i| < |Z_j|, Z_j > 0) + I(|Z_i| > |Z_j|, Z_i > 0). \tag{A.1.0.3}$$

[2] This derivation was mostly inspired by [14] and [15].

The double sum in (A.1.0.2) can be re-written by expanding $i \neq j$ into the two cases $i < j$ and $j < i$ as follows:

$$\sum_{i \neq j} I(Z_i > 0, |Z_j| < |Z_i|) = \sum_{i<j} I(Z_i > 0, |Z_j| < |Z_i|) + \sum_{j<i} I(Z_i > 0, |Z_j| < |Z_i|).$$

Switch the letters i and j in the right-most sum to get:

$$\sum_{i \neq j} I(Z_i > 0, |Z_j| < |Z_i|) = \sum_{i<j} I(Z_i > 0, |Z_j| < |Z_i|) + \sum_{i<j} I(Z_j > 0, |Z_i| < |Z_j|).$$

Then combine the two sums on the RHS since they are both summing over $i < j$:

$$\sum_{i \neq j} I(Z_i > 0, |Z_j| < |Z_i|) = \sum_{i<j} \left(I(Z_i > 0, |Z_j| < |Z_i|) + I(Z_j > 0, |Z_i| < |Z_j|) \right).$$

But we showed in (A.1.0.3) that the summand on the RHS is equal to $I(Z_i + Z_j > 0)$, so combining (A.1.0.2) and (A.1.0.3) yields the following formulation:

$$W_n^+ = \sum_{i=1}^{n} I(Z_i > 0) + \sum_{i<j} I(Z_i + Z_j > 0). \tag{A.1.0.4}$$

A.1.1 Distribution in the No-Ties Case

Since W_n^+ is a linear combination of two U-statistics, we can use this fact to derive its distribution. The first component of W_n^+ is based on the kernel, $h(z) = I(z > 0)$. The U-statistic itself is $U_1(Z_1, \ldots, Z_n) = \frac{1}{n} \sum_{i=1}^{n} I(Z_i > 0)$, and is the U-statistic used for the sign test. The second U-statistic is based on the kernel, $h(z_1, z_2) = I(z_1 + z_2 > 0)$, and the corresponding U-statistic is $U_2(Z_1, \ldots, Z_n) = \binom{n}{2}^{-1} \sum_{i<j} I(Z_i + Z_j > 0)$. Thus,

$$W_n^+ = nU_1(Z_1, \ldots, Z_n) + \binom{n}{2} U_2(Z_1, \ldots, Z_n). \tag{A.1.1.1}$$

If the mean of W_n^+ is subtracted from W_n^+, and if an appropriate sample-size scale factor is multiplied by the difference, we can then use (A.1.1.1) to obtain:

$$\begin{aligned} \frac{\sqrt{n}}{\binom{n}{2}} [W_n^+ - E(W_n^+)] &= \frac{\sqrt{n}}{\binom{n}{2}} n(U_1 - E(U_1)) + \frac{\sqrt{n}}{\binom{n}{2}} \binom{n}{2} (U_2 - E(U_2)) \\ &= \frac{n^{3/2}}{\binom{n}{2}} (U_1 - E(U_1)) + \sqrt{n}(U_2 - E(U_2)) \end{aligned}$$

By Corollary 3.2.5 in Randles and Wolfe [54], U_1 converges in quadratic mean (and hence in probability) to $E(U_1)$, so $U_1 - E(U_1)$ converges to 0 in probability. $\frac{n^{3/2}}{\binom{n}{2}}$ also converges to 0, so by Slutsky's Theorem,

$$\frac{\sqrt{n}}{\binom{n}{2}}[W_n^+ - E(W_n^+)] \text{ and } \sqrt{n}(U_2 - E(U_2))$$

have the same limiting distribution. Theorem 3.3.13 (One-Sample U-Statistic Theorem) in Randles and Wolfe [54] is then applicable to $\sqrt{n}(U_2 - E(U_2))$ as long as the following two conditions hold:

1. $E[h^2(Z_1, \ldots, Z_r)] < \infty$

2. $\zeta_1 = E[h(Z_1, Z_2, \ldots, Z_r)h(Z_1, Z_{r+1}, \ldots, Z_{2r-1})] - \gamma^2$ is positive.

Here, $\gamma = E[h(Z_1, Z_2, \ldots, Z_r)]$ is the parameter to be estimated, and the value of r is 2. For U_2, $h(Z_1, Z_2) = 1$ if $Z_1 + Z_2 > 0$ and zero otherwise, so $h^2(Z_1, Z_2) = h(Z_1, Z_2)$. Hence, $0 \leq E[h^2(Z_1, Z_2)] \leq 1$, and the first condition is satisfied. To show the other condition ($\zeta_1 > 0$), we begin with the definition of ζ_1 :

$$\zeta_1 = E[h(Z_1, Z_2)h(Z_1, Z_3)] - \{E[h(Z_1, Z_2)]\}^2$$

By applying repeated conditioning to both terms, this becomes:

$$\zeta_1 = E\{E[h(Z_1, Z_2)h(Z_1, Z_3)|Z_1]\} - \{E[E[h(Z_1, Z_2)|Z_1]]\}^2$$

Since $h(Z_1, Z_2)$ and $h(Z_1, Z_3)$ are conditionally independent given Z_1, we get the following:

$$\zeta_1 = E\{E[h(Z_1, Z_2)|Z_1] \cdot E[h(Z_1, Z_3)|Z_1]\} - \{E[E[h(Z_1, Z_2)|Z_1]]\}^2$$

Now since Z_i are assumed to be *i.i.d.* the following is true:

$$E[h(Z_1, Z_2)|Z_1] = E[h(Z_1, Z_3)|Z_1]$$

If we let $\Pi(h|Z_1) = E[h(Z_1, Z_2)|Z_1]$, then we get:

$$\zeta_1 = E[\Pi(h|Z_1)^2] - \{E[\Pi(h|Z_1)]\}^2.$$

This is the standard formula for the variance of X, $\mathrm{Var}(X) = E[X^2] - \{E[X]\}^2$. Hence,

$$\zeta_1 = Var(\Pi(h|Z_1)),$$

and since ζ_1 is the variance of a random variable, it is automatically non-negative. All that is left to show is $\zeta_1 \neq 0$, and this would be equivalent to the random variable $\Pi(h|Z_1)$ taking on more than one value with positive probability.

$$\Pi(h|Z_1) = E[h(Z_1, Z_2)|Z_1] = E[I(Z_1 + Z_2 > 0)|Z_1] = P[Z_1 + Z_2 > 0|Z_1],$$

and this probability will certainly vary as long as Z_1 takes more than one value. Hence ζ_1 is positive and the second condition of the One-Sample U-Statistic Theorem is satisfied. By applying the theorem to $\sqrt{n}(U_2 - E(U_2))$, we can conclude that $\sqrt{n}(U_2 - E(U_2))$ has a limiting normal distribution with mean 0 and variance $r^2\zeta_1 = 4\zeta_1$.

A.2 The Case Where Ties are Possible

The above derivation changes in a couple of ways when ties are present in the data. The type of ties considered here are when $|Z_i| = |Z_j|$ for $i \neq j$. This would mean that $|X_i - Y_i| = |X_j - Y_j|$. When the probability of this event is nonzero, we adjust the derivation above in the following two ways. First of all, equation (A.1.0.3) will no longer hold since now $|Z_i| = |Z_j|$ with positive probability. That equation now becomes:

$$\begin{aligned} I(Z_i + Z_j > 0) = \;& I(|Z_i| < |Z_j|, Z_j > 0) \\ +\;& I(|Z_i| > |Z_j|, Z_i > 0) \\ +\;& I(|Z_i| = |Z_j|, Z_j > 0, Z_i > 0). \end{aligned} \tag{A.2.0.2}$$

The previous formula for R_i^+, given by $R_i^+ = 1 + \sum_{j=1}^n I(|Z_j| < |Z_i|)$, will no longer hold since now we also need to count the number of Z_j that have exactly the same magnitude as Z_i. In order to create the tied rank, we'll average the ranks of any tied values, and replace these ranks by the average. The new formula for R_i^+ that works with or without ties is:

$$R_i^+ = \frac{1}{2} + \sum_{j=1}^n I(|Z_j| < |Z_i|) + \frac{1}{2}\sum_{j=1}^n I(|Z_j| = |Z_i|).$$

Putting this all together, we can begin to derive an appropriate formula for W_n^+ that works in both cases as well:

$$W_n^+ = \sum_{i=1}^n \left(\frac{1}{2} + \sum_{j=1}^n I(|Z_j| < |Z_i|) + \frac{1}{2}\sum_{j=1}^n I(|Z_j| = |Z_i|) \right) I(Z_i > 0).$$

Using the distributive law, this becomes:

$$\begin{aligned} W_n^+ = \;& \frac{1}{2}\sum_{i=1}^n I(Z_i > 0) \\ +\;& \sum_{i=1}^n \sum_{j=1}^n I(|Z_j| < |Z_i|) I(Z_i > 0) \\ +\;& \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n I(|Z_j| = |Z_i|) I(Z_i > 0). \end{aligned} \tag{A.2.0.3}$$

The second term on the RHS was reduced in the no-ties case (comparing equations (A.1.0.2) and (A.1.0.4)) to:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} I(|Z_j| < |Z_i|)I(Z_i > 0) = \sum_{i<j} I(Z_i + Z_j > 0),$$

but it will be slightly more complicated this time due to the additional term in equation (A.2.0.2). It is still the case that

$$\sum_{i=1}^{n}\sum_{j=1}^{n} I(|Z_j| < |Z_i|)I(Z_i > 0) = \sum_{i<j} \left(I(Z_i > 0, |Z_j| < |Z_i|) + I(Z_j > 0, |Z_i| < |Z_j|)\right),$$

but now by equation (A.2.0.2), we have

$$\sum_{i=1}^{n}\sum_{j=1}^{n} I(|Z_j| < |Z_i|)I(Z_i > 0) = \sum_{i<j} I(Z_i + Z_j > 0) - \sum_{i<j} I(|Z_i| = |Z_j|, Z_j > 0, Z_i > 0).$$

Substituting into equation (A.2.0.3), we get:

$$\begin{aligned} W_n^+ = & \frac{1}{2}\sum_{i=1}^{n} I(Z_i > 0) \\ & + \sum_{i<j} I(Z_i + Z_j > 0) - \sum_{i<j} I(|Z_i| = |Z_j|, Z_j > 0, Z_i > 0) \\ & + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} I(|Z_j| = |Z_i|)I(Z_i > 0). \end{aligned} \tag{A.2.0.4}$$

Let's examine this last term for a moment. Ignoring the $\frac{1}{2}$, the double sum can be split into two cases, according to whether $i = j$ or not:

$$\begin{aligned} \sum_{i=1}^{n}\sum_{j=1}^{n} I(|Z_j| = |Z_i|)I(Z_i > 0) = & \sum_{i=j} I(|Z_j| = |Z_i|)I(Z_i > 0) \\ & + \sum_{i\neq j} I(|Z_j| = |Z_i|)I(Z_i > 0). \end{aligned} \tag{A.2.0.5}$$

In the case where $i = j$, we are really just summing over i, and we are summing $I(Z_i > 0)$. The sum over $i \neq j$ can be broken into two parts based on $i < j$ and $j < i$.

$$\begin{aligned} \sum_{i=1}^{n}\sum_{j=1}^{n} I(|Z_j| = |Z_i|)I(Z_i > 0) = & \sum_{i=1}^{n} I(Z_i > 0) \\ & + \sum_{i<j} I(|Z_j| = |Z_i|)I(Z_i > 0) \\ & + \sum_{j<i} I(|Z_j| = |Z_i|)I(Z_i > 0). \end{aligned} \tag{A.2.0.6}$$

Switching the i and j in the final term on the RHS gives:

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n} I(|Z_j| = |Z_i|)I(Z_i > 0) = & \sum_{i=1}^{n} I(Z_i > 0) \\ & + \sum_{i<j} I(|Z_j| = |Z_i|)I(Z_i > 0) \\ & + \sum_{i<j} I(|Z_i| = |Z_j|)I(Z_j > 0).\end{aligned} \tag{A.2.0.7}$$

Combining the last two terms on the RHS then gives:

$$\begin{aligned}&\sum_{i=1}^{n}\sum_{j=1}^{n} I(|Z_j| = |Z_i|)I(Z_i > 0) = \\ &\qquad \sum_{i=1}^{n} I(Z_i > 0) + \sum_{i<j} I(|Z_j| = |Z_i|)I(Z_i > 0) + I(|Z_i| = |Z_j|)I(Z_j > 0).\end{aligned} \tag{A.2.0.8}$$

If $I(|Z_j| = |Z_i|)$ is factored out of the summand in the last term on the RHS, we are left with a value we'll call δ :

$$\delta = I(Z_i > 0) + I(Z_j > 0),$$

which can be 0, 1, or 2, depending on the signs of Z_i and Z_j. Hence, the sum can be split along those three cases:

$$\sum_{i<j} I(|Z_j| = |Z_i|)\delta = \sum_{\substack{i<j \\ \delta=2}} I(|Z_j| = |Z_i|)\ \delta + \sum_{\substack{i<j \\ \delta=1}} I(|Z_j| = |Z_i|)\ \delta + \sum_{\substack{i<j \\ \delta=0}} I(|Z_j| = |Z_i|)\ \delta, \tag{A.2.0.9}$$

which can then be re-written as (note that every term in the $\delta = 0$ sum was zero):

$$\sum_{i<j} \delta I(|Z_j| = |Z_i|) = \sum_{\substack{i<j \\ \delta=2}} 2I(|Z_j| = |Z_i|) + \sum_{\substack{i<j \\ \delta=1}} I(|Z_j| = |Z_i|). \tag{A.2.0.10}$$

The case of $\delta = 1$ implies that exactly one of Z_i, Z_j is positive and one is negative. Combining this with $|Z_j| = |Z_i|$ is equivalent to $Z_i + Z_j = 0$. Putting this all together we have:

$$\begin{aligned}\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} I(|Z_j| = |Z_i|)I(Z_i > 0) = {} & \frac{1}{2}\sum_{i=1}^{n} I(Z_i > 0) \\ & + \frac{1}{2}\sum_{\substack{i<j \\ \delta=2}} 2I(|Z_j| = |Z_i|) \\ & + \frac{1}{2}\sum_{i<j} I(Z_i + Z_j = 0).\end{aligned} \tag{A.2.0.11}$$

Substituting back into equation (A.2.0.4), we have:

$$\begin{aligned} W_n^+ = {} & \frac{1}{2}\sum_{i=1}^{n} I(Z_i > 0) + \sum_{i<j} I(Z_i + Z_j > 0) \\ & - \sum_{i<j} I(|Z_i| = |Z_j|, Z_j > 0, Z_i > 0) + \frac{1}{2}\sum_{i=1}^{n} I(Z_i > 0) \\ & + \sum_{\substack{i<j \\ \delta=2}} I(|Z_j| = |Z_i|) + \frac{1}{2}\sum_{i<j} I(Z_i + Z_j = 0).\end{aligned} \tag{A.2.0.12}$$

The third term and the fifth term on the RHS cancel out (since $\delta = 2$ means $Z_i > 0$ and $Z_j > 0$) and the first and fourth terms can be combined to give:

$$W_n^+ = \sum_{i=1}^{n} I(Z_i > 0) + \sum_{i<j} I(Z_i + Z_j > 0) + \frac{1}{2}\sum_{i<j} I(Z_i + Z_j = 0). \tag{A.2.0.13}$$

This new formula for W_n^+ actually works in both the case where ties are possible and the case where they are not. This is because when ties are not possible, $P(Z_i + Z_j = 0)$, which causes the third term to drop out. Note that the analysis done in Section A.1.1 can be adjusted slightly to work in both cases as well. The $h(z_1, z_2)$ used there can simply become $h(z_1, z_2) = I(Z_i + Z_j > 0) + \frac{1}{2}I(Z_i + Z_j = 0)$, and virtually every line of the analysis done in Section A.1.1 still holds. Hence, W_n^+ will be consistent against all alternatives for which $P(Z_i + Z_j > 0) > P(Z_i + Z_j < 0)$, for all i, j.

Appendix B

Annotated Model Summaries

The categories and aspects for TAC 2011 were developed based on model summaries from prior summarization tasks. Examples of these model summaries which have been annotated with the aspects are as follows:

- <aid="4.1">Ice</aid="4.1"> <aid="4.3">continues to melt at an alarming rate</aid="4.3"> <aid="4.1">in both the Arctic and Antarctic</aid="4.1">. <aid="4.3">Higher temperatures have shrunk the Arctic ice area 10% and its thickness 42% in 30 years</aid="4.3">. <aid="4.3">The permafrost is shrinking, endangering</aid="4.3"> <aid="4.2">infrastructure</aid="4.2">. <aid="4.3">These changes are threatening</aid="4.3"> <aid="4.2">the culture and economy of the indigenous Artic population</aid="4.2">. <aid="4.1">Ice shelves in the Antarctic</aid="4.1"> <aid="4.3">are collapsing</aid="4.3">. The <aid="4.3">melting</aid="4.3"> of the <aid="4.1">West Antarctic Ice Sheet</aid="4.1"> <aid="4.2">could raise ocean levels worldwide approximately 15 feet</aid="4.2">. <aid="4.3">Increased tourism in the Antarctic is having an environmental impact</aid="4.3">. <aid="4.4">Researchers are debating whether greenhouse gases or natural climate cycles are the biggest cause of the melting.</aid="4.4">

- <aid="4.3">Collapse of <aid="4.1">coastal Antarctic ice shelves</aid="4.1"> accelerated eight-fold the seaward flow of inland glaciers</aid="4.3">, <aid="4.2">raising sea levels</aid="4.2">: <aid="4.1">Larsen A (1995), Wilkins (1998), Larsen B (2002), Larsen C (this century)</aid="4.1">. <aid="4.3">Currents undermine <aid="4.1">the Ross and Ronne ice shelves</aid="4.1">, enabling ice flows from deep within the West Antarctic ice sheet.</aid="4.3"> <aid="4.1">Arctic permafrost</aid="4.1"> <aid="4.3">thawed</aid="4.3">; <aid="4.1">glaciers and sea ice</aid="4.1"> <aid="4.3">retreated</aid="4.3">. <aid="4.3">In 30 years <aid="4.1">the Arctic ice cap's</aid="4.1"> area shrank by 10%, its thickness by 42%,</aid="4.3"> <aid="4.2">opening shorter maritime routes when Arctic sea ice disappears in future summers</aid="4.2">. <aid="4.1">Siberian lakes</aid="4.1"> <aid="4.3">disappeared</aid="4.3">. <aid="4.2">Indigenous cultures and glacier tourism suffered</aid="4.2">. <aid="4.2">Bird migrations shifted</aid="4.2">. <aid="4.2">Northern Hemisphere weather will worsen.</aid="4.2">

- <aid="4.1">In Antarctica and the Artic, ice</aid="4.1"> <aid="4.3">melts</aid="4.3"> are causing complex questions about the impact of global warming. <aid="4.1">In Antarctica huge glaciers</aid="4.1"> <aid="4.3">are thinning</aid="4.3"> and <aid="4.1">ice shelves</aid="4.1"> <aid="4.3">are either disintegrating or retreating</aid="4.3">. These findings are possible indications of global warming. Information gathered about Antarctica coincides with a recent report on <aid="4.3">accelerating climate changes in the Arctic</aid="4.3">. A Chinese scientist predicted that <aid="4.1">the Artic icecap</aid="4.1"> <aid="4.3">would melt by 2080</aid="4.3">. <aid="4.2">The Arctic's indigenous people (about 4 million) <aid="4.4">are fighting global warming</aid="4.4"> because it will be a threat to their societies, economies and culture.</aid="4.2">

- <aid="4.3">The thinning of</aid="4.3"> <aid="4.1">glaciers and ice shelves</aid="4.1">, as well as <aid="4.3">the softening</aid="4.3"> <aid="4.1">of the permafrost</aid="4.1">, <aid="4.3">has accelerated greatly in recent years</aid="4.3">. While it is not certain that the man-made greenhouse effect is entirely to blame, <aid="4.4">it is clear that man must take steps now to address the problem</aid="4.4">. <aid="4.2">Global warming affects everything: oil-platforms, the society of peoples who are indiginous to the polar regions, polar animals, migratory birds, lakes (which are drying up as the permafrost melts), and even tourism.</aid="4.2"> <aid="4.2">As melting cold fresh water enters the salty sea, it will affect ocean currents and therefore world climate.</aid="4.2">

Bibliography

[1] Enrique Amigó, Julio Gonzalo, and Felisa Verdejo. The heterogeneity principle in evaluation measures for automatic summarization. In *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*, pages 36–43, Montréal, Canada, June 2012. Association for Computational Linguistics.

[2] Peter J. Bickel and Jian-Jian Ren. The Bootstrap in Hypothesis Testing. In *State of the Art in Statistics and Probability Theory, Festschrift for Willem R. van Zwet*, volume 36 of *Lecture Notes– Monograph Series*, pages 91–112. Institute of Mathematical Statistics, 2001.

[3] Ben Carterette and James Allan. Semiautomatic evaluation of retrieval systems using document similarities. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, pages 873–876, New York, NY, USA, 2007. ACM.

[4] William Jay Conover. On methods of handling ties in the wilcoxon signed-rank test. *Journal of the American Statistical Association*, 68(344):pp. 985–988, 1973.

[5] John M. Conroy and Hoa Trang Dang. Mind the Gap: Dangers of Divorcing Evaluations of Summary Content from Linguistic Quality. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 145–152, Manchester, UK, August 2008.

[6] John M. Conroy and Judith D. Schlesinger. CLASSY and TAC 2008 Metrics. In *TAC 2008 Workshop Proceedings*, 2008. http://www.nist.gov/tac/publications/index.html.

[7] John M. Conroy, Judith D. Schlesinger, and Dianne P. O'Leary. Nouveau-ROUGE: A Novelty Metric for Update Summarization. *Computational Linguistics*, 37(1):1–8, 2011.

[8] John M. Conroy, Judith D. Schlesinger, Peter A. Rankel, and Dianne P. O'Leary. Guiding CLASSY Toward More Responsive Summaries. In *TAC 2010 Workshop Proceedings*, 2010. http://www.nist.gov/tac/publications/index.html.

[9] Hoa T. Dang and Karolina Owczarzak. Overview of the TAC 2008 update summarization task. In *Proceedings of the 1st Text Analysis Conference (TAC)*, Gaithersburg, Maryland, USA, 2008.

[10] Hoa Trang Dang. Overview of DUC 2006. In *DUC 2006 Workshop Proceedings*, 2006. http://duc.nist.gov/pubs.html.

[11] Paulo C F de Oliveira, Edson Wilson Torrens, Alexandre Cidral, Sidney Schossland, and Evandro Bittencourt. Evaluating summaries automatically - a system proposal. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

[12] *Document Understanding Conference.* NIST, 2004. http://duc.nist.gov.

[13] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap.* Chapman & Hall, New York, 1993.

[14] Thomas S. Ferguson. Solutions for the Section on U-Statistics. http://www.math.ucla.edu/~tom/Stat200C/solUstat.pdf. Accessed: 2013-05-22.

[15] Thomas S. Ferguson. U-Statistcs: Notes for Statistics 200C, Spring 2005. `http://www.math.ucla.edu/~tom/Stat200C/Ustat.pdf`. Accessed: 2013-05-22.

[16] Deen G. Freelon. ReCal: Intercoder reliability calculation as a web service. *International Journal of Internet Science*, 5(1), 2010.

[17] George Giannakopoulos and Vangelis Karkaletsis. Autosummeng and memog in evaluating guided summaries. In *Proceedings of the Text Analysis Conference (TAC 2011)*, Gaithersburg, Maryland, USA, 2011. NIST.

[18] George Giannakopoulos, Vangelis Karkaletsis, George Vouros, and Panagiotis Stamatopoulos. Summarization system evaluation revisited: N-gram graphs. *ACM Trans. Speech Lang. Process.*, 5(3):5:1–5:39, October 2008.

[19] George Giannakopoulos, George A. Vouros, and Vangelis Karkaletsis. Mudos-ng: Multi-document summaries using n-gram graphs (tech report). *CoRR*, abs/1012.2042, 2010.

[20] Yvette Graham. Re-evaluating automatic summarization with bleu and 192 shades of rouge. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 128–137, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[21] Yvette Graham, Nitika Mathur, and Timothy Baldwin. Randomized significance tests in machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 266–274, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics.

[22] J. Hájek and Z. Šidák. *Theory of Rank Tests*. Czechoslovak Academy of science. Academic Press, 1967.

[23] M. Hollander and D.A. Wolfe. *Nonparametric Statistical Methods*. Wiley series in probability and statistics: Texts and references section. Wiley, 1999.

[24] Eduard Hovy, Chin-Yew Lin, Liang Zhou, and Junichi Fukumoto. Automated summarization evaluation with basic elements. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 899–902, 2006.

[25] David Hull. Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '93, pages 329–338, New York, NY, USA, 1993. ACM.

[26] D. Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.

[27] Tapas Kanungo and David Orr. Predicting the readability of short web summaries. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 202–211, New York, NY, USA, 2009. ACM.

[28] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

[29] Alistair Kennedy, Anna Kazantseva Saif Mohammad, Terry Copeck, Diana Inkpen, and Stan Szpakowicz. Getting emotional about news. In *Fourth Text Analysis Conference (TAC 2011)*, 2011.

[30] Klaus Krippendorff. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70, 1970.

[31] Niraj Kumar, Kannan Srinathan, and Vasudeva Varma. Using unsupervised system with least linguistic features for tac-aesop task. In *Fourth Text Analysis Conference (TAC 2011)*, 2011.

[32] Niraj Kumar, Kannan Srinathan, and Vasudeva Varma. Using graph based mapping of co-occurring words and closeness centrality score for summarization evaluation.

In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 7182 of *Lecture Notes in Computer Science*, pages 353–365. Springer Berlin / Heidelberg, 2012.

[33] Erich Leo Lehmann and H.J.M. D'Abrera. *Nonparametrics: statistical methods based on ranks*. Holden-Day series in probability and statistics. Holden-Day, 1975.

[34] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In Stan Szpakowicz Marie-Francine Moens, editor, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[35] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In Stan Szpakowicz Marie-Francine Moens, editor, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[36] Annie Louis and Ani Nenkova. Automatically evaluating content selection in summarization without human models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 306–314, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[37] Annie Louis and Ani Nenkova. Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, 39:267–300, 2013.

[38] H P Luhn. The automatic creation of literature abstracts. In *Advances in Automatic Text Summarization*, pages 58–63. The MIT Press, 1956.

[39] Kathleen McKeown and Dragomir R. Radev. Generating summaries of multiple news articles. In Edward A. Fox, Peter Ingwersen, and Raya Fidel, editors, *SIGIR*, pages 74–82. ACM Press, 1995.

[40] Ani Nenkova. Discourse factors in multi-document summarization. In *AAAI*, pages 1654–1655, 2005.

[41] Ani Nenkova and Annie Louis. Can you summarize this? identifying correlates of input difficulty for multi-document summarization. In *ACL*, pages 825–833, 2008.

[42] Ani Nenkova and Rebecca Passonneau. Evaluating content selection in summarization: The pyramid method. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 145–152, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics.

[43] Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Trans. Speech Lang. Process.*, 4(2), May 2007.

[44] G.E. Noether. *Elements of nonparametric statistics*. SIAM series in applied mathematics. Wiley, 1967.

[45] P. Over. Introduction to DUC-2001: an intrinsic evaluation of generic news text summarization systems. Technical report, Retrieval Group, Information Access Division, National Institute of Standards and Technology, 2001.

[46] Karolina Owczarzak, John M. Conroy, Hoa Trang Dang, and Ani Nenkova. An assessment of the accuracy of automatic evaluation in summarization. In *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*, pages 1–9, Montréal, Canada, June 2012. Association for Computational Linguistics.

[47] Karolina Owczarzak, Peter A. Rankel, Hoa Trang Dang, and John M. Conroy. Assessing the effect of inconsistent assessors on summarization evaluation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 359–362, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

[48] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[49] Rebecca J. Passonneau, Ani Nenkova, Kathleen McKeown, and Sergey Sigelman. Applying the Pyramid method in DUC 2005. In *Proceedings of the 5th Document Understanding Conference (DUC)*, Vancouver, Canada, 2005.

[50] Emily Pitler, Annie Louis, and Ani Nenkova. Automatic evaluation of linguistic quality in multi-document summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 544–554, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[51] John W. Pratt. Remarks on zeros and ties in the wilcoxon signed rank procedures. *Journal of the American Statistical Association*, 54(287):pp. 655–667, 1959.

[52] Joseph Putter. The treatment of ties in some nonparametric tests. *The Annals of Mathematical Statistics*, 26(3):pp. 368–386, 1955.

[53] *Pyramid Annotation Guide: DUC 2006*. Columbia University, 2006. `http://www1.cs.columbia.edu/~becky/DUC2006/2006-pyramid-guidelines.html`.

[54] R.H. Randles and D.A. Wolfe. *Introduction to the Theory of Nonparametric Statistics*. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. Wiley, 1979.

[55] Peter Rankel, John Conroy, Eric Slud, and Dianne O'Leary. Ranking human and machine summarization systems. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 467–473, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

[56] Peter A. Rankel, John M. Conroy, Hoa Trang Dang, and Ani Nenkova. A decade of automatic content evaluation of news summaries: Reassessing the state of the art. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 131–136, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[57] Peter A. Rankel, John M. Conroy, and Judith D. Schlesinger. Better metrics to automatically predict the quality of a text summary. *Algorithms*, 5(4):398–420, 2012.

[58] Horacio Saggion, Juan-Manuel Torres-Moreno, Iria da Cunha, and Eric SanJuan. Multilingual summarization evaluation without human models. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 1059–1067, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[59] Mark Sanderson, Falk Scholer, and Andrew Turpin. Relatively relevant: Assessor shift in document judgements. In *Proceedings of the Australasian Document Computing Symposium*, pages 60–67, 2010.

[60] Mark Sanderson and Ian Soboroff. Problems with kendall's tau. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 839–840, New York, NY, USA, 2007. ACM.

[61] Henry Scheffé. *The Analysis of Variance*. Wiley publication in mathematical statistics. John Wiley & Sons, 1959.

[62] Falk Scholer, Andrew Turpin, and Mark Sanderson. Quantifying test collection quality based on the consistency of relevance judgements. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 1063–1072, New York, NY, USA, 2011. ACM.

[63] G.A.F. Seber. *Multivariate observations.* Wiley series in probability and statistics. Wiley-Interscience, 2004.

[64] Mark D. Smucker, James Allan, and Ben Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 623–632, New York, NY, USA, 2007. ACM.

[65] Mark D. Smucker, James Allan, and Ben Carterette. Agreement among statistical significance tests for information retrieval evaluation at varying sample sizes. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 630–631, New York, NY, USA, 2009. ACM.

[66] J. Steinberger and K. Ježek. Evaluation measures for text summarization. *Computing and Informatics*, 28(2):251–275, 2012.

[67] Tac 2008 update summary task. `http://www.nist.gov/tac/2008/summarization/update_summarization.instructions.pdf`, 2008.

[68] *Text Analysis Conference.* NIST, 2011. `http://www.nist.gov/tac`.

[69] J. Tavernier and P. Bellot. Combining relevance and readability for inex 2011 question-answering track. In *Pre-Proceedings of INEX 2011*, pages 185–195, Amsterdam, 2011. IR Publications.

[70] Stephen Tratz and Eduard Hovy. Summarisation evaluation using transformed basic elements. In *Proceedings TAC 2008*. NIST, 2008.

[71] Alan Turing. Computing Machinery and Intelligence. *Mind*, 59(236):433–460, 1950.

[72] András Vargha and Harold D. Delaney. The kruskal-wallis test and stochastic homogeneity. *Journal of Educational and Behavioral Statistics*, 23(2):pp. 170–192, 1998.

[73] John von Neumann and Oskar Morgenstern. *Theory of games and economic behavior.* Princeton Univ. Press, Princeton, NJ, 3. ed. edition, 1953.

[74] Wikipedia. Krippendorff's alpha — wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=Krippendorff%27s_alpha&oldid=713891629`, 2016.

[75] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):pp. 80–83, 1945.

[76] Frank Wilcoxon. Some rapid approximate statistical procedures. *Annals of the New York Academy of Sciences*, 52(6):808–814, 1950.

[77] Alexander Yeh. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th conference on Computational linguistics - Volume 2*, COLING '00, pages 947–953, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.

[78] Emine Yilmaz, Javed A. Aslam, and Stephen Robertson. A new rank correlation coefficient for information retrieval. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 587–594, New York, NY, USA, 2008. ACM.