Extractive vs. NLG-based Abstractive Summarization of Evaluative Text: The Effect of Corpus Controversiality

Giuseppe Carenini and Jackie Chi Kit Cheung[1]
Department of Computer Science
University of British Columbia
Vancouver, B.C. V6T 1Z4, Canada
{carenini,cckitpw}@cs.ubc.ca

Abstract

Extractive summarization is the strategy of concatenating extracts taken from a corpus into a summary, while abstractive summarization involves paraphrasing the corpus using novel sentences. We define a novel measure of corpus controversiality of opinions contained in evaluative text, and report the results of a user study comparing extractive and NLG-based abstractive summarization at different levels of controversiality. While the abstractive summarizer performs better overall, the results suggest that the margin by which abstraction outperforms extraction is greater when controversiality is high, providing a context in which the need for generation-based methods is especially great.

1 Introduction

There are two main approaches to the task of summarization—extraction and abstraction (Hahn and Mani, 2000). Extraction involves concatenating extracts taken from the corpus into a summary, whereas abstraction involves generating novel sentences from information extracted from the corpus. It has been observed that in the context of multi-document summarization of news articles, extraction may be inappropriate because it may produce summaries which are overly verbose or biased towards some sources (Barzilay et al., 1999). However, there has been little work identifying specific factors which might affect the performance of each strategy in summarizing evaluative documents containing opinions and preferences, such as customer reviews or blogs. This work aims to address this gap by exploring one dimension along which the effectiveness of the two paradigms could vary; namely, the controversiality of the opinions contained in the corpus.

In this paper, we make the following contributions. Firstly, we define a measure of controversiality of opinions in the corpus based on information entropy. Secondly, we run a user study to test the hypothesis that a controversial corpus has greater need of abstractive methods and consequently of NLG techniques. Intuitively, extracting sentences from multiple users whose opinions are diverse and wide-ranging may not reflect the overall opinion, whereas it may be adequate content-wise if opinions are roughly the same across users. As a secondary contribution, we propose a method for structuring text when summarizing controversial corpora. This method is used in our study for generating abstractive summaries.

The results of the user study support our hypothesis that a NLG summarizer outperforms an extractive summarizer by more when the controversiality is high.

2 Related Work

There has been little work comparing extractive and abstractive multi-document summarization. A previous study on summarizing evaluative text (Carenini et. al, 2006) showed that extraction and abstraction performed about equally well, though for different reasons. The study, however, did not

[1]Authors are listed in alphabetical order.

look at the effect of the controversiality of the corpus on the relative performance of the two strategies.

To the best of our knowledge, the task of measuring the controversiality of opinions in a corpus has not been studied before. Some well known measures are related to this task, including variance, information entropy, and measures of inter-rater reliability. (e.g. Fleiss' Kappa (Fleiss, 1971), Krippendorff's Alpha (Krippendorff, 1980)). However, these existing measures do not satisfy certain properties that a sound measure of controversiality should possess, prompting us to develop our own based on information entropy.

Summary evaluation is a challenging open research area. Existing methods include soliciting human judgements, task-based approaches, and automatic approaches.

Task-based evaluation measures the effectiveness of a summarizer for its intended purpose. (e.g. (McKeown et al., 2005)) This approach, however, is less applicable in this work because we are interested in evaluating specific properties of the summary such as the grammaticality and the content, which may be difficult to evaluate with an overall task-based approach. Furthermore, the design of the task may intrinsically favour abstractive or extractive summarization. As an extreme example, asking for a list of specific comments from users would clearly favour extractive summarization.

Another method for summary evaluation is the Pyramid method (Nenkova and Passonneau, 2004), which takes into account the fact that human summaries with different content can be equally informative. Multiple human summaries are taken to be models, and chunks of meaning known as Summary Content Units (SCU) are manually identified. Peer summaries are evaluated based on how many SCUs they share with the model summaries, and the number of model summaries in which these SCUs are found. Although this method has been tested in DUC 2006 and DUC 2005 (Passonneau et al., 2006), (Passonneau et al., 2005) in the domain of news articles, it has not been tested for evaluative text. A pilot study that we conducted on a set of customer reviews on a product using the Pyramid method revealed several problems specific to the evaluative domain. For example, summaries which misrepresented the polarity of the evaluations for a certain feature were not penalized, and human summaries sometimes produced contradictory statements about the distribution of the opinions. In one case, one model summary claimed that a feature is positively rated, while another claimed the opposite, whereas the machine summary indicated that this feature drew mixed reviews. Clearly, only one of these positions should be regarded as correct. Further work is needed to resolve these problems.

There are also automatic methods for summary evaluation, such as ROUGE (Lin, 2004), which gives a score based on the similarity in the sequences of words between a human-written model summary and the machine summary. While ROUGE scores have been shown to often correlate quite well with human judgements (Nenkova et al., 2007), they do not provide insights into the specific strengths and weaknesses of the summary.

The method of summarization evaluation used in this work is to ask users to complete a questionnaire about summaries that they are presented with. The questionnaire consists of questions asking for Likert ratings and is adapted from the questionnaire in (Carenini et al., 2006).

3 Representative Systems

In our user study, we compare an abstractive and an extractive multi-document summarizer that are both developed specifically for the evaluative domain. These summarizers have been found to produce quantitatively similar results, and both significantly outperform a baseline summarizer, which is the MEAD summarization framework with all options set to the default (Radev et al., 2000).

Both summarizers rely on information extraction from the corpus. First, sentences with opinions need to be identified, along with the features of the entity that are evaluated, the strength, and polarity (positive or negative) of the evaluation. For instance, in a corpus of customer reviews, the sentence “Excellent picture quality - on par with my Pioneer, Panasonic, and JVC players.” contains an opinion on the feature *picture quality* of a DVD player, and is a very positive evaluation (+3 on a scale from -3 to +3). We rely on methods from previous work for these tasks (Hu and Liu, 2004). Once these features, called Crude Features (CFs), are extracted, they are mapped onto a taxonomy of User Defined Features (UDFs), so named because they can be defined by the user. This mapping provides a better conceptual organization of the CFs

by grouping together semantically similar CFs, such as *jpeg picture* and *jpeg slide show* under the UDF *JPEG*. For the purposes of our study, feature extraction, polarity/strength identification and the mapping from CFs to UDFs are not done automatically as in (Hu and Liu, 2004) and (Carenini et al, 2005). Instead, "gold standard" annotations by humans are used in order to focus on the effect of the summarization strategy.

3.1 Abstractive Summarizer: SEA

The abstractive summarizer is the Summarizer of Evaluative Arguments (SEA), adapted from GEA, a system for generating evaluative text tailored to the user's preferences (Carenini and Moore, 2006).

In SEA, units of content are organized by UDFs. The importance of each UDF is based on the number and strength of evaluations of CFs mapped to this UDF, as well as the importance of its children UDFs. Content selection consists of repeating the following two steps until the desired number of UDFs have been selected: (i) greedily selecting the most important UDF (ii) recalculating the measure of importance scores for the remaining UDFs.

The content structuring, microplanning, and realization stages of SEA are adapted from GEA. Each selected UDF is realized in the final summary by one clause, generated from a template pattern based on the number and distribution of polarity/strength evaluations of the UDF. For example, the UDF *video output* with an average polarity/strength of near -3 might be realized as "several customers found the video output to be terrible."

While experimenting with the SEA summarizer, we noticed that the document structuring of SEA summaries, which is adapted from GEA and is based on guidelines from argumentation theory (Carenini and Moore, 2000), sometimes sounded unnatural. We found that controversially rated UDF features (roughly balanced positive and negative evaluations) were treated as contrasts to those which were uncontroversially rated (either mostly positive, or mostly negative evaluations). In SEA, contrast relations between features are realized by cue phrases signalling contrast such as "however" and "although". These cue phrases appear to signal a contrast that is too strong for the relation between controversial and uncontroversial features. An example of a SEA summary suffering from this problem can be found in Figure 1.

> Customers had mixed opinions about the Apex AD2600. Although several customers found the video output to be poor and some customers disliked the user interface, customers had mixed opinions about the range of compatible disc formats. However, users did agree on some things. Some users found the extra features to be very good even though customers had mixed opinions about the supplied universal remote control.

Figure 1: SEA summary of a controversial corpus with a document structuring problem. Controversial and uncontroversial features are interwoven. See Figure 3 for an example of a summary structured with our alternative strategy.

To solve this problem, we devised an alternative content structure for controversial corpora, in which all controversial features appear first, followed by all positively and negatively evaluated features.

3.2 Extractive Summarizer: MEAD*

The extractive approach is represented by MEAD*, which is adapted from the open source summarization framework MEAD (Radev et al., 2000).

After information extraction, MEAD* orders CFs by the number of sentences evaluating that CF, and selects a sentence from each CF until the word limit has been reached. The sentence that is selected for each CF is the one with the highest sum of polarity/strength evaluations for any feature, so sentences that mention more CFs tend to be selected. The selected sentences are then ordered according to the UDF hierarchy by a depth-first traversal through the UDF tree so that more abstract features tend to precede more specific ones.

MEAD* does not have a special mechanism to deal with controversial features. It is not clear how overall controversiality of a feature can be effectively expressed with extraction, as each sentence conveys a specific and unique opinion. One could include two sentences of opposite polarity for each controversial feature. However, in several cases that we considered, this produced extremely incoherent text that did not seem to convey the gist of the overall controversiality of the feature.

3.3 Links to the Corpus

In common with the previous study on which this is based, both the SEA and MEAD* summaries contain "clickable footnotes" which are links back into an original user review, with a relevant sentence highlighted. These footnotes serve to provide details for the abstractive SEA summarizer, and context for the sentences chosen by the extractive MEAD* summarizer. They also aid the participants of the user study in checking the contents of the summary. The sample sentences for SEA are selected by a method similar to the MEAD* sentence selection algorithm. One of the questions in the questionnaire provided to users targets the effectiveness of the footnotes as an aid to the summary.

4 Measuring Controversiality

The opinion sentences in the corpus are annotated with the CF that they evaluate as well as the strength, from 1 to 3, and polarity, positive or negative, of the evaluation. It is natural then, to base a measure of controversiality on these annotations. To measure the controversiality of a corpus, we first measure the controversiality of each of the features in the corpus. We list two properties that a measure of feature controversiality should satisfy.

Strength-sensitivity: The measure should be sensitive to the strength of the evaluations. e.g. Polarity/strength (P/S) evaluations of -2 and +2 should be less controversial than -3 and +3

Polarity-sensitivity: The measure should be sensitive the polarity of the evaluations. e.g. P/S evaluations of -1 and +1 should be more controversial than +1 and +3.

The rationale for this property is that positive and negative evaluations are fundamentally different, and this distinction is more important than the difference in intensity. Thus, though a numerical scale would suggest that -1 and +1 are as distant as +1 and +3, a suitable controversiality measure should not treat them so.

In addition, the overall measure of corpus controversiality should also satisfy the following two features.

CF-weighting: CFs should be weighted by the number of evaluations they contain when calculating the overall value of controversiality for the corpus.

CF-independence: The controversiality of individual CFs should not affect each other.

An alternative is to calculate controversiality by UDFs instead of CFs. However, not all CFs mapped to the same UDF represent the same concept. For example, the CFs *picture clarity* and *color signal* are both mapped to the UDF *video output*.

4.1 Existing Measures of Variability

Since the problem of measuring the variability of a distribution has been well studied, we first examined existing metrics including variance, entropy, kappa, weighted kappa, Krippendorff's alpha, and information entropy. Each of these, however, is problematic in their canonical form, leading us to devise a new metric based on information entropy which satisfies the above properties. Existing metrics will now be examined in turn.

Variance: Variance does not satisfy polarity-sensitivity, as the statistic only takes into account the difference of each data point to the mean, and the sign of the data point plays no role.

Information Entropy: The canonical form of information entropy does not satisfy strength or polarity sensitivity, because the measure considers the discrete values of the distribution to be an unordered set.

Measures of Inter-rater Reliability: Many measures exist to assess inter-rater agreement or disagreement, which is the task of measuring how similarly two or more judges rate one or more subjects beyond chance (dis)agreement. Various versions of Kappa and Krippendorff's Alpha (Krippendorff, 1980), which have shown to be equivalent in their most generalized forms (Passonneau, 1997), can be modified to satisfy all the properties listed above. However, there are important differences between the tasks of measuring controversiality and measuring inter-rater reliability. Kappa and Krippendorff's Alpha correct for chance agreement between raters, which is appropriate in the context of inter-rater reliability calculations, because judges are asked to give their opinions on items that are given to them. In contrast, expressions of opinion are volunteered by users, and users self-select the features they comment on. Thus, it is reasonable to assume that they never randomly select an evaluation for a feature, and chance agreement does not exist.

4.2 Entropy-based Controversiality

We define here our novel measure of controversiality, which is based on information entropy because it can be more easily adapted to measure controversiality. As has been stated, entropy in its original form over the evaluations of a CF is not sensitive to strength or polarity. To correct this, we first aggregate the positive and negative evaluations for each CF separately, and then calculate the entropy based on the resultant Bernoulli distribution.

Let $ps(cf_j)$ be the set of polarity/strength evaluations for cf_j. Let the importance of a feature, $imp(cf_j)$, be the sum of the absolute values of the polarity/strength evaluations for cf_j.

$$imp(cf_j) = \sum_{ps_k \in ps(cf_j)} |ps_k|$$

Define:

$$imp\_pos(cf_j) = \sum_{ps_k \in ps(cf_j) \wedge ps_k > 0} |ps_k|$$

$$imp\_neg(cf_j) = \sum_{ps_k \in ps(cf_j) \wedge ps_k < 0} |ps_k|$$

Now, calculate the entropy of the Bernoulli distribution corresponding to the importance of the two polarities to satisfy polarity-sensitivity. That is, Bernoulli with parameter

$$\theta_j = imp\_pos(cf_j) / imp(cf_j)$$

$$H(\theta_j) = -\theta_j \log_2 \theta_j - (1-\theta_j)\log_2(1-\theta_j)$$

Next, we scale this score by the importance of the evaluations divided by the maximum possible importance for this number of evaluations to satisfy strength-sensitivity. Since our scale is from -3 to +3, the maximum possible importance for a feature is three times the number of evaluations.

$$max\_imp(cf_j) = 3 \times |ps(cf_j)|$$

Then the controversiality of a feature is:

$$contro(cf_j) = \frac{imp(cf_j) \times H(\theta_j)}{max\_imp(cf_j)}$$

The case corresponding to the highest possible feature controversiality, then, would be the bimodal case with equal numbers of evaluations on the extreme positive and negative bins (Figure 2). Note, however, that controversiality is not simply bimodality. A unimodal normal-like distribution of evaluations centred on zero, for example, should intuitively be somewhat controversial, because there are equal numbers of positive and negative evaluations. Our entropy-based feature controversiality measure is able to take this into account.

To calculate the controversiality of the corpus, a weighted average is taken over the CF controversiality scores, with the weight being equal to one less than the number of evaluations for that CF. We subtract one to eliminate any CF where only one evaluation is made, as that CF has an entropy score of one by default before scaling by importance. This procedure satisfies properties CF-weighting and CF-independence.

$$w(cf_j) = |ps(cf_j)| - 1$$

$$contro(corpus) = \frac{\sum w(cf_j) \times contro(cf_j)}{\sum w(cf_j)}$$

Although the annotations in this corpus range from -3 to +3, it would be easy to rescale opinion annotations of different corpora to apply this metric. Note that empirically, this measure correlates highly with Kappa and Krippendorff's Alpha.

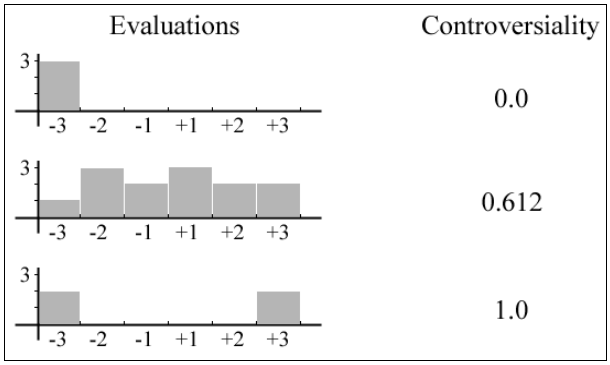

Figure 2: Sample feature controversiality scores for three different distributions of polarity/strength evaluations.

5 User Study

Our main hypothesis that extractive summarization is outperformed even more in the case of controversial corpora was tested by a user study, which compared the results of MEAD* and the modified SEA. First, ten subsets of 30 user reviews were selected from the corpus of 101 reviews of the Apex AD2600 DVD player from amazon.com by stochastic local search. Five of these subsets are controversial, with controversiality scores between 0.83 and 0.88, and five of these are uncontroversial, with controversiality scores of 0. A set of thir-

SEA	MEAD*
Customers had mixed opinions about the Apex AD2600 1,2 possibly because users were divided on the range of compatible disc formats 3,4 and there was disagreement among the users about the video output 5,6. However, users did agree on some things. Some purchasers found the extra features 7 to be very good and some customers really liked the surround sound support 8 and thought the user interface 9 was poor.	When we tried to hook up the first one , it was broken - the motor would not eject discs or close the door . 1 The build quality feels solid , it does n't shake or whine while playing discs , and the picture and sound is top notch (both dts and dd5.1 sound good) . 2 The progressive scan option can be turned off easily by a button on the remote control which is one of the simplest and easiest remote controls i have ever seen or used . 3 It plays original dvds and cds and plays mp3s and jpegs . 4

Figure 3: Sample SEA and MEAD* summaries for a controversial corpus. The numbers within the summaries are footnotes linking the summary to an original user review from the corpus.

ty user reviews per subcorpus was needed to create a summary of sufficient length, which in our case was about 80 words in length.

Twenty university students were recruited and presented with two summaries of the same subcorpus, one generated from SEA and one from MEAD*. We generated ten subcorpora in total, so each subcorpus was assigned to two participants. One of these participants was shown the SEA summary first, and the other was shown the MEAD* summary first, to eliminate the order of presentation as a source of variation.

The participants were asked to take on the role of an employee of Apex, and told that they would have to write a summary for the quality assurance department of the company about the product in question. The purpose of this was to prime them to look for information that should be included in a summary of this corpus. They were given thirty minutes to read the reviews, and take notes.

They were then presented with a questionnaire on the summaries, consisting of ten Likert rating questions. Five of these questions targeted the linguistic quality of the summary, based on linguistic well-formedness questions used at DUC 2005, one targeted the "clickable footnotes" linking to sample sentences in the summary (see section 3.3), and three evaluated the contents of the summary. The three questions targeted *Recall*, *Precision*, and the general *Accuracy* of the summary contents respectively. The tenth question asked for a general overall quality judgement of the summary.

After familiarizing themselves with the questionnaire, the participants were presented with the two summaries in sequence, and asked to fill out the questionnaire while reading the summary. They were allowed to return to the original set of reviews during this time. Lastly, they were given an additional questionnaire which asked them to compare the two summaries that they were shown. Questions in the questionnaire not found in (Carenini et al., 2006) are attached in Appendix A.

6 Results

6.1 Quantitative Results

We convert the Likert responses from a scale from Strongly Disagree to Strong Agree to a scale from 1 to 5, with 1 corresponding to Strongly Disagree, and 5 to Strongly Agree. We group the ten questions into four categories: linguistic (questions 1 to 5), content (questions 6 to 8), footnote (question 9), and overall (question 10). See Table 1 for a breakdown of the responses for each question at each controversiality level.

For our analysis, we adopt a two-step approach that has been applied in Computational Linguistics (Di Eugenio et al., 2002) as well as in HCI (Hinckley et al., 1997).

First, we perform a two-way Analysis of Variance (ANOVA) test using the average response of the questions in each category. The two factors are controversiality of the corpus (high or low) as independent samples, and the summarizer (SEA or MEAD*) as repeated measures. We repeat this procedure for the average of the ten questions, termed *Macro* below. The p-values of these tests are summarized in Table 2.

The results of the ANOVA tests indicate that SEA significantly outperforms MEAD* in terms of linguistic and overall quality, as well as for all the questions combined. It does not significantly outperform MEAD* by content, or in the amount that the included sample sentences linked to by the footnotes aid the summary. No significant differences are found in the performance of the summa-

	Controversial						Uncontroversial					
	SEA		MEAD*		(SEA – MEAD*		SEA		MEAD*		(SEA – MEAD*	
Question	Mean	St. Dev.	Mean	St. Dev.	Mean	St. Dev.	Mean	St. Dev.	Mean	St. Dev.	Mean	St. Dev.
Grammaticality	4.5	0.53	3.4	1.26	1.1	0.99	4.2	0.92	2.78	1.3	1.56	1.51
Non-redundancy	4.2	0.92	4	1.07	0.25	1.58	3.7	0.95	3.8	1.14	-0.1	1.45
Referential clarity	4.5	0.53	3.44	1.33	1	1.22	4.2	1.03	3.5	1.18	0.7	1.34
Focus	4.11	1.27	2.1	0.88	2.22	0.83	3.9	1.1	2.6	1.35	1.3	1.57
Structure and Coherence	4.1	0.99	1.9	0.99	2.2	1.14	3.8	1.4	2.3	1.06	1.5	1.9
Linguistic	*4.29*	*0.87*	*2.91*	*1.35*	*1.39*	*1.34*	*3.96*	*1.07*	*3*	*1.29*	*0.98*	*1.63*
Recall	2.8	1.32	1.8	1.23	1	1.33	2.5	1.27	2.5	1.43	0	1.89
Precision	3.9	1.1	2.7	1.64	1.2	1.23	3.5	1.27	3.3	0.95	0.2	1.93
Accuracy	3.4	0.97	3.3	1.57	0.1	1.2	3.1	1.52	3.2	1.03	-0.1	2.28
Content	*3.37*	*1.19*	*2.6*	*1.57*	*0.77*	*1.3*	*3.03*	*1.38*	*3*	*1.17*	*0.03*	*1.97*
Footnote	4	1.05	3.9	0.88	0.1	1.66	3.6	1.07	3.5	1.35	0.1	1.6
Overall	3.8	0.79	2.4	1.17	1.4	1.07	3.2	1.23	2.7	0.82	0.5	1.84
Macro – Footnote	*3.92*	*1.06*	*2.75*	*1.41*	*1.17*	*1.32*	*3.57*	*1.26*	*2.97*	*1.2*	*0.61*	*1.81*
Macro	*3.93*	*1.05*	*2.87*	*1.4*	*1.06*	*1.39*	*3.57*	*1.24*	*3.02*	*1.22*	*0.56*	*1.79*

Table 1: Breakdown of average Likert question responses for each summary at the two levels of controversiality as well as the difference between SEA and MEAD*.

Question Set	Controversiality	Summarizer	Controversiality x Summarizer
Linguistic	0.7226	<0.0001	0.2639
Content	0.9215	0.1906	0.2277
Footnote	0.2457	0.7805	1
Overall	0.6301	0.0115	0.2000
Macro	*0.7127*	*0.0003*	*0.1655*

Table 2: Two-way ANOVA p-values.

rizers over the two levels of controversiality for any of the question sets .

While the average differences in scores between the SEA and MEAD* summarizers are greater in the controversial case for the linguistic, content, and macro averages as well as the question on the overall quality, the p-values for interaction between the two factors in the two-way ANOVA test are not significant.

For the second step of the analysis, we use a one-tailed sign test (Siegel and Castellan, 1988) over the difference in performance of the summarizers at the two levels of controversiality for the questions in the questionnaire. We encode + in the case where the difference between SEA and MEAD* is greater for a question in the controversial setting, – if the difference is smaller, and we discard a question if the difference is the same (e.g. the *Footnote* question). Since the *Overall* question is likely correlated with the responses of the other questions, we did not include it in the test. After discarding the *Footnote* question, the p-value over the remaining eight questions is 0.0352, which lends support to our hypothesis that the abstraction is better by more when the corpus is controversial.

We also analyze the users' summary preferences at the two levels of controversiality. A strong preference for SEA is encoded as a 5, while a strong preference for MEAD* is encoded as a 1, with 3 being neutral. Using a two-tailed unpaired two-sample t-test, we do not find a significant difference in the participants' summary preferences (p=0.6237). However, participants sometimes preferred summaries for reasons other than linguistic or content quality, or may base their judgement only on one aspect of the summary. For instance, one participant rated SEA at least as well as MEAD* in all questions except *Footnote*, yet preferred MEAD* to SEA overall precisely because MEAD* was felt to have made better use of the footnotes than SEA.

6.2 Qualitative Results

The qualitative comments that participants were asked to provide along with the Likert scores confirmed the observations that led us to formulate the initial hypothesis.

In the controversial subcorpora, participants generally agreed that the abstractive nature of SEA's generated text was an advantage. For example, one participant lauded SEA for attempting to "synthesize the reviews" and said that it "did reflect the mixed nature of the reviews, and covered some common complaints." The participant, however, said that SEA "was somewhat misleading in that it understated the extent to which reviews were negative. In particular, agreement was reported on

some features where none existed, and problems with reliability were not mentioned."

Participants disagreed on the information coverage of the MEAD* summary. One participant said that MEAD* includes "almost all the information about the Apex 2600 DVD player", while another said that it "does not reflect all information from the customer reviews."

In the uncontroversial subcorpora, more users criticized SEA for its inaccuracy in content selection. One participant felt that SEA "made generalizations that were not precise or accurate." Participants had specific comments about the features that SEA mentioned that they did not consider important. For example, one comment was that "Compatibility with CDs was not a general problem, nor were issues with the remote control, or video output (when it worked)." MEAD* was criticized for being "overly specific", but users praised MEAD* for being "not at all redundant", and said that it "included information I felt was important."

7 Conclusion and Future Work

We have explored the controversiality of opinions in a corpus of evaluative text as an aspect which may determine how well abstractive and extractive summarization strategies perform. We have presented a novel measure of controversiality, and reported on the results of a user study which suggest that abstraction by NLG outperforms extraction by a larger amount in more controversial corpora. We have also presented a document structuring strategy for summarization of controversial corpora.

Our work has implications in practical decisions on summarization strategy choice; an extractive approach, which may be easier to implement because of its lack of requirement for natural language generation, may suffice if the controversiality of opinions in a corpus is sufficiently low.

A future approach to summarization of evaluative text might combine extraction and abstraction in order to combine the different strengths that each bring to the summary. The controversiality of the corpus might be one factor determining the mix of abstraction and extraction in the summary. The footnotes linking to sample sentences in the corpus in SEA are already one form of this combined approach. Further work is needed to integrate this text into the summary itself, possibly in a modified form.

As a final note to our user study, further studies should be done with different corpora and summarization systems to increase the external validity of our results.

Acknowledgements

We would like to thank Raymond T. Ng, Gabriel Murray and Lucas Rizoli for their helpful comments. This work was partially funded by the Natural Sciences and Engineering Research Council of Canada.

References

Regina Barzilay, Kathleen R. McKeown, and Michael Elhadad. 1999. Information fusion in the context of multi-document summarization. In *Proc. 37th ACL*, pages 550–557.

Giuseppe Carenini and Johanna D. Moore. 2006. Generating and evaluating evaluative arguments. *Artificial Intelligence*, 170(11):925-952.

Giuseppe Carenini, Raymond Ng and Adam Pauls. 2006. Multi-document summarization of evaluative text. In *Proc. 11th EACL 2006*, pages 305-312.

Giuseppe Carenini, Raymond T. Ng and Ed Zwart. 2005. Extracting Knowledge from Evaluative Text. In *Proc. 3rd International Conference on Knowledge Capture*, pages 11-18.

Giuseppe Carenini and Johanna D. Moore. 2000. A strategy for generating evaluative arguments. In *First INLG*, pages 47-54, Mitzpe Ramon, Israel.

Barbara Di Eugenio, Michael Glass, and Michael J. Trolio. 2002. The DIAG experiments: Natural language generation for intelligent tutoring systems. In *INLG02, The 2nd INLG*, pages 120-127.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*. 76:378-382.

U. Hahn and I. Mani. 2000. The challenges of automatic summarization. *IEEE Computer*, 33(11):29-36.

Ken Hinckley, Randy Pausch, Dennis Proffitt, James Patten, and Neal Kassell. 1997. Cooperative bimanual action. In *Proc. CHI Conference*.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proc. 10th ACM SIGKDD conference*, pages 168-177.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Beverly Hills, CA.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proc. of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004*, Barcelona, Spain.

2005. Linguistic quality questions from the 2005 document understanding conference. http://duc.nist.gov/duc2005/quality-questions.txt

Kathleen McKeown, Rebecca Passonneau, David Elson, Ani Nenkova, and Julia Hirschberg. 2005. Do summaries help? A task-based evaluation of multi-document summarization. In *Proc. SIGIR 2005*.

Ani Nenkova, Rebecca J. Passonneau, and K. McKeown. 2007. The pyramid method: incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing*, 4(2).

Ani Nenkova and Rebecca J. Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proc. NAACL/HLT*.

Rebecca J. Passonneau, Kathleen McKeown, Sergey Sigleman, and Adam Goodkind. 2006. Applying the pyramid method in the 2006 Document Understanding Conference. In *Proc. DUC'06*.

Rebecca J. Passonneau, Ani Nenkova, Kathleen McKeown, and Sergey Sigleman. 2005. Applying the pyramid method in DUC 2005. In *Proc. DUC'05*.

Rebecca J. Passonneau. 1997. Applying Reliability Metrics to Co-Reference Annotation. Department of Computer Science, Columbia University, TR CUCS-017-97.

Dragomir Radev, Hongyan Jing, and Malgorzata Budzikowska. 2000. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *Proc. ANLP/NAACL Workshop on Automatic Summarization*.

S. Siegel and N. J. Castellan, Jr. 1988. Nonparametric statistics for the behaviorial sciences. McGraw Hill.

Appendix A. Additional Questions

Footnotes: a) Did you use the footnotes when reviewing the summary?

b) Answer this question only if you answered "Yes" to the previous question. The clickable footnotes were a helpful addition to the summary.

Summary Comparison Questions:

1) List any Pros and Cons you can think of for each of the summaries. Point form is okay.

2) Overall, which summary did you prefer?

3) Why did you prefer this summary? (If the reason overlaps with some points from question 1, put a star next to those points in the chart.)

4) Do you have any other comments about the reviews or summaries, the tasks, or the experiment in general? If so, please write them below.