Summarization of Evaluative Text:

The effect of corpus controversiality on extractive and abstractive summarization

Jackie CK Cheung
University of British Columbia
cckitpw@cs.ubc.ca

Abstract

In automatic summarization, 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. In this paper, we define a novel measure of corpus controversiality, and report the results of a user study comparing extractive and abstractive summarization at different levels of controversiality. While the abstractive summarizer performs better overall, the results suggest that margin by which abstraction outperforms extraction may be greater when controversiality is high, though the difference was not found in this study to be statistically significant.

1 Introduction

One recent trend in online media is an increasing interactivity of websites. An example of this is that many websites now allow users to provide feedback on commercial products that they have purchased and used. As the quantity of these reviews grow, so too does the need for automatic summarization methods for them.

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 paraphrasing the corpus using novel sentences. It has been observed that in the context of multi-document summarization, such as summarizing customer reviews, extraction may be inappropriate because it might 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, especially when summarizing evaluative text containing opinions and preferences. 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 summary corpus.

The hypothesis that is tested is that a controversial corpus has greater need of abstractive methods, because extracting sentences from multiple users whose opinions are diverse and wide-ranging might not reflect the overall opinion. Conversely, extracting sentences might be adequate for summary content if opinions are roughly the same across users.

To measure the controversiality of the opinions in a corpus, we define a measure of controversiality of opinions in the corpus based on information entropy. We conducted a user study to examine the effectiveness of two summarization systems, one representing each strategy, on corpora of differing levels of controversiality.

The corpora from which the summaries are generated are subsets of user reviews of the the AD2600 DVD player taken from amazon.com. User opinions in the reviews have been annotated with strength and polarity of evaluation (positive or negative) (Hu and Liu, 2004). Based on these polarity/strength values, the subsets are selected to have either a high or low level of controversiality.

The results of the user study somewhat support the above hypothesis, as the margin by which abstraction outperforms extraction is larger when controversiality is high. Although this difference is not statistically significant, qualitative comments by study participants suggest that a larger sample size would have yielded statistical significance.

2 Related Work

2.1 Comparing Extraction and Abstraction

There has been little work specifically comparing extractive and abstractive summarization. A previous study (Carenini et. al, 2006) with the same summarization systems showed that extraction and abstraction performed about equally well, though for different reasons. The study, however, did not look at the effect of the controversiality of the corpus on the relative performance of the two summarizers.

2.2 Measures of Distribution

Measuring the controversiality of opinions in a corpus is equivalent to measuring the spread of a distribution over the positive and negative evaluations. There are many existing measures for this purpose, such as variance, and information entropy, which measures the uncertainty associated with a random variable. Measures of inter-rater reliability have also been developed, for example Cohen's Kappa (Cohen, 1960), Fleiss' Kappa (Fleiss, 1971), Weighted Kappa (Cohen, 1968), and Krippendorff's Alpha (Krippendorff, 1980). These existing measures do not satisfy certain properties that a measure of controversiality should have, prompting us to develop our own, based on information entropy. See section 4.2 for a detailed discussion.

2.3 Summarization Evaluation

The method of summarization evaluation used in this work is to ask users to fill out a questionnaire about the summaries that they are presented with. The questionnaire consists of questions asking for Likert ratings and is closely modelled after the questionnaire in (Carenini et al., 2006), with additional questions specifically asking participants to compare the two summaries that they see.

There exist more sophisticated methods of evaluation aside from directly soliciting judgements via a questionnaire. One is to take a task-based approach, which is to measure the effectiveness of the system for its intended purpose. For example, the effectiveness of the Generator of Evaluative Arguments evaluative text generator (Carenini and Moore, 2006) was determined by the degree to which it influenced the users in the task. This approach, however, is less applicable in this work because we are interested in specific properties of the summary such as the grammaticality and the content, and the individual effects of these properties may be difficult to detect with an overall task-based approach. Furthermore, the design of the task may intrinsically favour abstractive or extractive summarization. For example, asking users to give the overall opinion from the summaries may inherently favour abstractive summarization, while asking them to give specific examples of criticisms that customers made may favour extractive summarization.

Another method for summary evaluation is the Pyramid method (Nenkova and Passonneau, 2004), which takes into account the fact that multiple 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.

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. The principle reason that this method is not suitable for our purposes is that it has not been used in the evaluative domain. A pilot study that we conducted using the Pyramid method highlighted several problems in applying the method directly to the evaluative domain. For example, summaries which misrepresented the polarity of the evaluations for a certain feature are not penalized, and human summaries sometimes produced contradictory statements about the distribution of the opinions. For instance, one model summary claimed that a particular feature was positively rated, while another claimed the opposite, whereas the machine summary indicated that this feature drew mixed opinions from the users. Clearly, only one of these positions should be regarded as correct. Further work is needed to resolve this problem.

Furthermore, the Pyramid method is very labour intensive, as a new set of human generated summaries must be gathered for each corpus.

At the other end of the spectrum, there are automatic methods for summary evaluation, such as ROUGE (Lin, 2004), which give 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 correlate quite well with human judgements (Nenkova et al., 2007), the downside to this approach is that the scores do not provide any insights into the specific strengths and weaknesses of the summary.

3 Representative Systems

The representative summarization systems of extractive and abstractive summarization were both developed specifically for the evaluative domain (Carenini et al, 2006). 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).

Before	**After**
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.	Customers had mixed opinions about the Apex AD2600 possibly because users were divided on the range of compatible disc formats and customers had mixed opinions about the video output. However, users did agree on some things. Some customers really liked the extra features and some users thought the surround sound support was very good and disliked the user interface.

Figure 1: Sample SEA summaries of controversial corpora before and after modification to document structuring.

Opinion sentences, features and polarity/strength are identified and extracted by methods from previous work (Hu and Liu, 2004). One innovation of these summarization systems is that the surface-level Crude Features (CFs) of the evaluated entity that are extracted are mapped onto a hierarchical structure of User Defined Features (UDFs) (Carenini et al., 2005), so named because they are defined by the user, and can be user-tailored depending on the features that the user considers important in evaluating a product. This mapping is done by word similarity metrics, and it provides a better conceptual organization of the CFs by allowing CFs that are superficially different (such as "remote" and "remote control") to be treated as the same UDF ("remote control"). The tree hierarchy also provides a mechanism to model the relationships between features.

For the purposes of this study, feature extraction, polarity/strength identification and the mapping from CFs to UDFs are not done automatically. Instead, "gold standard" annotations by humans are used in order to focus on the effect of the summarization strategy itself.

3.1 Extractive Summarizer: MEAD*

The extractive approach is represented by MEAD*, which is adapted from the open source summarization framework MEAD.

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.

3.2 Abstractive Summarizer: SEA

The representative summarizer for the abstractive approach is the Summarizer of Evaluative Arguments (SEA), adapted from GEA.

SEA selects by UDFs, in contrast to the CF-based method of MEAD*. Each UDF has a measure of importance based on the number of strength of evaluations evaluating CFs mapped to this UDF, as well as the measure of importance of children UDFs which have not been selected yet. Feature selection consists of greedily selecting the UDF with the highest measure of importance and then recalculating the measure of importance scores for the 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 distribution of opinions and polarity/strength 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 was not very natural. The document structuring of SEA is adapted from GEA, and is based on guidelines from argumentation theory (Carenini and Moore, 2000).

The problem was that UDF features which are controversially rated, in other words, features with roughly equal proportions of positive and negative evaluations, are treated as contrasting features to UDF features which are positively or negatively rated. In SEA, contrast relations are realized by the document structuring and microplanning stages with cue phrases signally contrast such as "however" and "although". This, however, appears to be inappropriate, because these cue phrases signal a contrast that is too strong for the relation. An example of a SEA summary suffering from this problem can be found in Figure 1.

To avoid putting SEA at a disadvantage in controversial corpora, which contain many instances of controversial features, we implemented an alternative content structure for controversiality corpora, in which all controversial features appear first, followed by all positively and negatively evaluated features. This eliminates any instances of the above situation, and seems to result in a more coherent summary.

3.3 Sample Sentences

In common with the previous study on which this is based, both the SEA and MEAD* summaries contained “clickable footnotes” which serve as links back into the original set of user reviews. These footnotes serve to provide details for the abstractive SEA summarizer, and context to the extractive MEAD* summarizer. They also aid the participants of the user study in checking the contents of the summary. The sample sentences were selected by a similar method to the MEAD* sentence selection algorithm. One of the questions in the questionnaire that participants are presented with for each summary specifically asks for the effectiveness of the footnotes as an aid to the summary.

4 Measuring Controversiality

4.1 Properties of a Controversiality Measure

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 is on these annotations.

The following are various properties that we want our measure of controversiality to satisfy, along with a specific case for illustration where warranted.

1. *Ordinality*

The measure should handle ordinal data.

e.g. Evaluations with polarity/strength (P/S) values of -3 and +1 for a feature should be more controversial than a +2 and +3.

2. *Strength-sensitivity*

The measure should be sensitive to the strength of the evaluations.

e.g. P/S evaluations of -2 and +2 should be less controversial than -3 and +3

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.

4. *CF-weighting*

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

5. *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.”

4.2 Comparing Potential Measures of Controversiality

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

Variance

Variance does not satisfy polarity-sensitivity.

Information Entropy

The canonical form of information entropy does not satisfy ordinality or strength-sensitivity, but a modified version satisfies all five properties above, as will be shortly shown.

Measures of Inter-rater Reliability

Many measures exist to measure inter-rater agreement or disagreement beyond chance, which is the task of measuring how similarly two or more judges rate one or more subjects. For example, various versions of Kappa (eg. Cohen's (Cohen, 1960), Fleiss' (Fleiss, 1971), Weighted (Cohen, 1968)) differ to the generality of the cases they handle.

A more recently devised measure, Krippendorff's Alpha, is more general and handles any kind of scale, with any number of raters and subjects. Kappa has been shown to be equivalent to Krippendorff's Alpha in their most generalized forms (Passonneau, 1997).

While these metrics can be modified to satisfy all the properties listed above, it is important to note that measuring the controversiality of a corpus is not the same as measuring inter-rater reliability. Kappa and Krippendorff's Alpha correct for chance agreement, 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 the context of expressions of opinion, however, users volunteer their opinions of features, and can self-select which features they comment on. Thus, it is reasonable to assume that they never randomly select an evaluation of a feature, and correcting for chance agreement is a liability rather than an asset.

4.3 Entropy-based Controversiality

We define here our novel measure of controversiality, which is based on information entropy. As has been stated, entropy in its original form over the evaluations of a CF does not satisfy ordinality, nor is it sensitive to strength and polarity. To correct this, we first aggregate the positive and negative evaluations for each CF separately, and then calculate the information entropy based on the resultant distribution, which is a Bernoulli distribution.

Formally, let $ps(cf_j)$ be the set of polarity/strength evaluations for cf_j. Define

$$support(cf_j) = \sum_{ps_k \, \varepsilon \, ps(cf_j),\, ps_k>0} | ps_k |$$

$$opposition(cf_j) = \sum_{ps_k \, \varepsilon \, ps(cf_j),\, ps_k<0} | ps_k |$$

$$fervour(cf_j) = support(cf_j) + opposition(cf_j)$$

Calculate

$$H(Ber(support(cf_j)/fervour(cf_j)) = - support(cf_j)/fervour(cf_j) * \log_2 (support(cf_j)/fervour(cf_j)) - opposition(cf_j)/fervour(cf_j) * \log_2 (opposition(cf_j)/fervour(cf_j))$$

Next, we scale this score by the strength of the evaluations. We define the fervour to be the sum of the strengths of the evaluations for a CF. Since our scale is from -3 to +3, the maximum fervour for this number of evaluations is 3 * number of evaluations. The above binary entropy score is scaled by multiplying the CF's fervour divided by the maximum fervour.

$$max\_fervour(cf_j) = 3 * |ps(cf_j)|$$

$$controversiality(cf_j) = fervour(cf_j) / max\_fervour(cf_j) * H(Ber(support(cf_j)/fervour(cf_j)))$$

The above calculation is done for each CF. This results in a controversiality score for each CF. To calculate the controversiality of the corpus, a weighted average is taken over the individual CF 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 1 by default before scaling by fervour.

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

$$controversiality(corpus) = \sum (w(cf_j) * controversiality(cf_j)) / \sum w(cf_j)$$

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.

5 User Study

To test the above hypothesis, a user study was conducted to compare the results of the 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 a local search method. 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 thirty user reviews per subcorpus was needed to create a summary of sufficient length, which in our case was about 80 words in length. A larger set of reviews reduces the maximum controversiality score that the local search method could find in a subset of this size. Furthermore, it would be more onerous for the participants to read through all the reviews.

We originally planned to test another corpus of 43 reviews of the Canon G3 digital camera. However, the opinions in this corpus were mostly positive, so we were unable to generate subcorpora of high enough contro-

SEA	**MEAD***
Almost all customers really disliked the Apex AD2600 1. Although some users thought the surround sound support 2 was very good and several customers loved the user interface 3, several customers found the range of compatible disc formats 4 to be poor. Furthermore, several purchasers thought the video output 5 was very poor. However, there were some positive evaluations. Some purchasers really liked the extra features 6 even though several users thought the supplied universal remote control 7 was very poor.	This product sucks , the customer service from apex sucks . 1 It 's very sleek looking with a very good front panel button layout , and it has a great feature set . 2 Can 't say whether i rec 'd an " updated " model but it will not read dvd + rw 's or vcd 's for me . 3 The unit seems to play all formats that i have put in it (jpeg , kodak pic 's and dvd-r) i have read other reviews and some good and soom not so good , but my feeling at this time is " two thumbs up " ! 4

Figure 2: Sample SEA and MEAD* summaries for an uncontroversial corpus. The numbers within the summaries are footnotes linking the summary to an original user review from the corpus.

	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.

versiality. Since we would not be able to test this corpus at high and low levels of controversiality, inclusion of this corpus into the study would introduce the confounding variable of the product type. Thus, we decided to set aside this corpus in this test.

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, in order to eliminate the order of presentation as a source of variation.

The methodology of the user study is adapted from a previous study (Carenini et al., 2006). The participants were asked at first to pretend that they were 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 then given thirty minutes to read the reviews, and to take notes of any information they find noteworthy.

They were then presented with a questionnaire on the summaries, consisting of ten Likert rating questions. Five of these were questions targeting the linguistic quality of the summary, based on SEE 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.

A copy of the questionnaire is attached in the appendix. Figure 2 shows sample SEA and MEAD* summaries for one of the uncontroversial corpora.

6 Results

We now report the results of the user study. We will first report the quantitative results of the Likert questions in the questionnaires, and then select some illustrative qualitative comments that participants provided.

6.1 Quantitative Results

We converted 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 grouped 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.

Using the average response of the questions in each category, we performed a two-way Analysis of Variance (ANOVA) test. The two factors were controversiality of the corpus (high or low) as independent samples, and the summarizer used (SEA or MEAD*) as repeated measures. We also repeated this procedure for the overall average of the ten questions. The results of these tests are summarized in Table 2.

	Controversiality	Summarizer	Interaction
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.

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.

None of the tests indicate a significant difference in the performance of the summarizers at 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. Thus, our initial hypothesis that controversiality favours abstraction more than extraction is not supported statistically.

Finally, we look at the preferences that users had for between the two summarizers at the two levels of controversiality. A strong preference for SEA was encoded as a 5, while a strong preference for MEAD* was encoded as a 1, with 3 being neutral. Using a two-tailed unpaired two-sample t-test with the null hypothesis that the samples at the two levels of controversiality had the same mean, we did not find a significant difference in the participants' preferences for one summary over the other (p=0.6237). It should be noted, however, that participants sometimes preferred summaries for reasons other than linguistic or content quality, or may base their judgement only on one aspect of the summary, as their qualitative comments revealed.

6.2 Qualitative Results

Although the interaction between controversiality and summarizer type was not found to be statistically significant, this result may be due to our small sample size. The qualitative comments that participants were asked to provide along with the Likert scores included the same observations that led us to formulate the initial hypothesis.

In the controversial subcorpora, participants generally agreed that the abstractive nature of SEA 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."

This participant strongly preferred SEA to MEAD*, with the following comment about MEAD*: "Since the extracts were so specific and taken out of context, the resulting 'summary' read incoherently and did not represent the set of reviews."

Participants disagreed on the degree to which MEAD* reflected the information content of the user reviews, with one participant saying 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 than in the controversial cases. 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."

In general, many users reacted very negatively towards MEAD*'s extractive nature at both levels of controversiality. For instance, one participant's comment about MEAD*'s summarization strategy was that "even I can figure out how to do it."

Several users commented on the omission of customer service in the SEA summary. This is due to the UDF hierarchy, which did not include a node for customer service. In a stroke of luck for MEAD*, customer service is mentioned, because a sentence selected for another reason also happens to mention customer service: "The product sucks, the customer service from apex sucks."[1] Redefining the UDF hierarchy by adding such a node would solve this problem.

The additional questionnaire that asked participants to specifically compare the two summaries that they were shown included a question asking them which summary they preferred. While this served as a check against their answers to the overall quality of the two summaries, it also highlighted the important fact that not all users prefer a summary for reasons of content and linguistic quality.

For instance, one participant rated SEA at least as well as MEAD* in all questions except the question on the footnote, yet preferred MEAD* to SEA overall because MEAD* was felt to have made better use of the footnotes than SEA. This participant did add that "if summary B [SEA] had appropriate footnotes and include[d] price information, I would have been very happy with it."

[1] Participants did also comment that this sentence caused the summary to sound rather unprofessional.

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 presented a novel measure of controversiality, and reported on the results of a user study which suggest that abstraction may outperform extraction by a larger amount in more controversial corpora. This 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.

The qualitative comments from the user study suggest that the abstractive and extractive summaries performed well in terms of their content for different reasons. The abstractive summaries were praised because they were generalizations that synthesized the information in the original corpus, while the extractive summaries were praised because they were more accurate as they came directly from the user reviews.

A future approach to summarization might combine extraction and abstraction in order to combine the different strengths that each bring to the summary. The footnotes linking to sample sentences in the corpus in SEA are already one form of this combination approach. What now needs to be done is to integrate this text into the summary itself, possibly in a modified form.

Although the statistical results of the user study did not support our initial hypothesis to statistical significance, a future user study with a larger number of users may be able to rectify this problem.

References

R. Barzilay, K. R. McKeown, and M. Elhadad. 1999. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, pages 550–557.

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

G. Carenini, R. Ng and A. Pauls. 2006. Multi-document summarization of evaluative text. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics: EACL 2006*, pages 305-312.

G. Carenini, R. Ng and E. Zwart. 2005. Extracting knowledge from evaluative text. In *Proc. Third International Conference on Knowledge Capture.*

G. Carenini and J. D. Moore. 2000. A strategy for generating evaluative arguments. In *First International Conference on Natural Language Generation*, pages 47-54, Mitzpe Ramon, Israel.

J. Cohen. 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin.* 70:213-20.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement.* 20:37-46.

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

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

M. Hu, and B. Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD conference,* pages 168-177.

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

C-Y. Lin. 2004. *ROUGE: A Package for Automatic Evaluation of Summaries*. In *Proceedings 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

A. Nenkova, R. 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).

A. Nenkova, and R. J. Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics and Human Language Technology (NAACL/HLT).*

R. J. Passonneau, K. McKeown, S. Sigleman, and A. Goodkind. 2006. Applying the pyramid method in the 2006 Document Understanding Conference. In *Proceedings of the Document Understanding Conference (DUC'06).*

R. J. Passonneau, A. Nenkova, K. McKeown, and S. Sigleman. 2005. Applying the pyramid method in DUC 2005. In *Proceedings of the Document Understanding Conference (DUC'05).*

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

D Radev, H. Jing, and M. Budzikowska. 2000. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *Proceedings of the ANLP/NAACL 2000 Workshop on Automatic Summarization*.

Appendix

User Study Questionnaire

This is a questionnaire about the automatic summary you see before you. You may explain your answers in the "comments" section if you wish.

Select one choice for each question which best represents your opinion. Please tell the experimenter when you are done.

Remember to ask the experimenter if there is anything that you are unsure of.

1 *Grammaticality*
The summary has 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 is 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 is easy to identify who or what the pronouns and noun phrases in the summary are referring to. If a person or other entity is mentioned, it is clear what their role in the summary 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 has a focus; sentences only contain information that is related to the rest of the summary.

5 *Structure and Coherence*
The summary is well-structured and well-organized. The summary is not just a heap of related information, but builds from sentence to sentence to a coherent body of information about the product reviews.

6 *Recall*
The summary contains all of the information you would have included from the source text.

7 *Precision*
The summary contains no information you would NOT have included from the source text.

8 *Accuracy*
All information expressed in the summary accurately reflects the information contained in the source text.

9 *Footnotes*
9a. Did you use the footnotes when reviewing the summary?

9b. Answer this question only if you answered "Yes" to the previous question.
The clickable footnotes were a helpful addition to the summary.

10 *Overall*
Overall, this summary was a good summary.

Here are some additional questions specifically asking you to compare the two summaries you saw during this hour.

Remember to ask the experimenter if there is anything that you are unsure of.

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.