Comparing Abstractive and Extractive Summarization of Evaluative Text: Controversiality and Content Selection

Jackie CK Cheung

Submitted in partial fulfillment of the
requirements for the degree
of B. Sc. (Hons.)
in the Department of Computer Science
of the Faculty of Science

UNIVERSITY OF BRITISH COLUMBIA
2008

Abstract

One of the main aspects of the so-called “Web 2.0” is increased participation by website users, or a blurring of the distinction between the content provider and the content receiver. One form that this user interaction can take is the sharing of comments on products that users have purchased or services that they have used. Examples abound on websites such as amazon.com, flixster.com, and chapters.indigo.ca. The need for efficient and effective multi-document summarization of these user reviews and other kinds of evaluative text containing opinions and preferences is thus ever-growing.

This thesis examines two canonical strategies for summarization: summarization by extraction, which consists of concatenating source sentences into a summary, and summarization by abstraction, which involves generating novel sentences for the summary (Hahn and Mani, 2000). The first part of this thesis compares the two summarization strategies when they are applied to the domain of summarizing evaluative text (e.g. user reviews). We report on the results of a user study which examines the interaction of the summarization strategy with the controversiality of the opinions in the corpus. We then propose a clustering framework for summarization content selection that allows us to combine the two strategies in order to capitalize on the strengths of each. We apply this framework to the summarization of evaluative text, using a clustering paradigm from the field of location theory called the p-median problem (Resende and Werneck, 2004).

Contents

1. Introduction **1**

2. Controversiality and Summarization Strategy **2**

2.1. Abstractive and Extractive Summarization 2

2.2. Summarization Evaluation 3

2.3. Representative Systems 4

2.3.1. Abstractive Summarizer: SEA 5

2.3.2. Extractive Summarizer: MEAD* 6

2.3.3. Links to the Corpus 7

2.4. Measuring Corpus Controversiality 7

2.4.1. Existing Measures of Variability 8

2.4.2. Entropy-based Controversiality 9

2.5. User Study 11

2.6. Results 12

2.6.1. Quantitative Results 12

2.6.2. Qualitative Results 14

2.7. Discussion 15

3. P-Median for Feature and Strategy Selection **16**

3.1. Clustering As a Framework for Content Selection 16

3.2. P-Median for Clustering 17

3.2.1. Definition of P-Median 17

3.2.2. Related Problems 18

3.2.3. Computational Complexity of P-Median 19

3.3. Defining Information Coverage 19

3.3.1. P-Median for Feature Selection 20

3.3.2. Simultaneous Feature and Strategy Selection 24

3.3.3. Extraction as Complementing Abstraction 27

3.4. Discussion and Future Work 31

4. Conclusion 32

References 33

Appendices 36

Appendix A. User Study Questionnaire 36

Appendix B. Enforcing Abstraction Before Extraction 37

List of Figures

2.1. SEA summary with structuring problem 5

2.2. SEA summary showing clickable footnotes 7

2.3. SEA and MEAD* summaries of a controversial corpus 9

2.4. Feature controversiality scores for three sample distributions 11

3.1. Percent agreement between greedy and p-median feature selection 23

List of Tables

2.1. Average Likert ratings in user study 13

2.2. Two-way ANOVA p-values 13

Acknowledgements

As graduate school looms ever closer, I am reassured by thinking of all the people who have supported me and have placed their confidence in me, to whom I would like to show my appreciation.

I would first like to thank my supervisor, Dr. Giuseppe Carenini, who introduced me to the world of research and guided me in taking my first steps into the field of computational linguistics. He showed endless patience with me whenever I popped by his office with a random, semi-coherent idea. Equal thanks must go to Dr. Raymond T. Ng, who has been like a second supervisor, providing me with much guidance and support.

To Lucas Rizoli, Jan Ulrich, the members of the summarization group, and the other members of the Laboratory for Computational Intelligence at UBC, thank you for many fruitful and enlightening discussions. I have had a fantastic year, because of you.

I would like to thank my parents for their unwavering and unconditional love, and my older brother Vincent and older sister Joyce for being my role models and for laughing at my corny jokes. My family is such an important part of my life, that I cannot imagine a life without them.

And finally, a big thanks to all my friends, some of whom I have known for nearly all my life. I do not know who I would be without you.

1. Introduction

One of the main aspects of the so-called "Web 2.0" is increased participation by website users, or a blurring of the distinction between the content provider and the content receiver. One form that this user interaction can take is the sharing of comments on products that users have purchased or services that they have used. Examples abound on websites such as amazon.com, flixster.com, and chapters.indigo.ca. The need for efficient and effective multi-document summarization of these user reviews and other kinds of evaluative text containing opinions and preferences is thus ever-growing.

There are two canonical strategies for summarization: summarization by extraction, which consists of concatenating source sentences into a summary, and summarization by abstraction, which involves generating novel sentences for the summary (Hahn and Mani, 2000). In the first part of this thesis, we explore one aspect of the corpus which may influence the choice of which strategy to employ; namely, the controversiality of opinions. We define a novel measure of corpus controversiality, and report on the results of a user study we ran to compare the two strategies at different levels of controversiality. The results support our hypothesis that abstraction outperforms extraction by a greater margin when corpus controversiality is high. In other words, the need for abstraction is especially high when opinions are diverse.

Having identified controversiality as one factor that might affect summary strategy choice, we attempt to use this knowledge to combine extraction and abstraction in order to capitalize on the strengths of each. We propose a clustering framework for summary content selection which is also able to select the summarization strategy that expresses the selected content. We apply this framework to summarizing evaluative text, using a well-studied clustering problem from the field of facility location theory called the p-median problem. The goal of this problem is to minimize the total distance from "customers" needing service to selected "facilities" which serve them. We suggest methods of applying this problem to content and strategy selection, and ways to evaluate them.

The contributions of this thesis are of interest to the summarization community in the following ways. Firstly, our definition of corpus controversiality and its influence on the effectiveness on extractive and abstractive summarization are important for practical decision making for applications where summarization is needed. If it is known that the controversiality of opinions is low enough, such as in the domain of news articles, the benefit of abstractive summarization, if any, may not be great enough to justify the added difficulty inherent with generating novel sentences (Carenini and Cheung, 2008). Secondly, our work on combining extraction and abstraction represents a shift away from the purely extractive approach which has been dominant in the past. This is a necessary shift, as pure extraction is proving inadequate for multi-document summarization (Barzilay et al., 1999), and especially for evaluative text, where opinions can be diverse and contradictory.

2. Controversiality and Summarization Strategy

2.1. Abstractive and Extractive Summarization

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 chapter 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.

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 by showing that a NLG summarizer outperforms an extractive summarizer to a larger extent when the controversiality is high.

2.2. Summarization Evaluation

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 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), which was itself based on linguistic well-formedness questions used at DUC 2005.

2.3. 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. The first step is the identification of sentences containing opinions, the features of the entity that are evaluated, and 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.

2.3.1. 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).

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 2.1: SEA summary of a controversial corpus with a document structuring problem. Controversial and uncontroversial features are interwoven. See Figure 2.3 for an example of a summary structured with our alternative strategy.

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 2.1.

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.

2.3.2. Extractive Summarizer: MEAD*

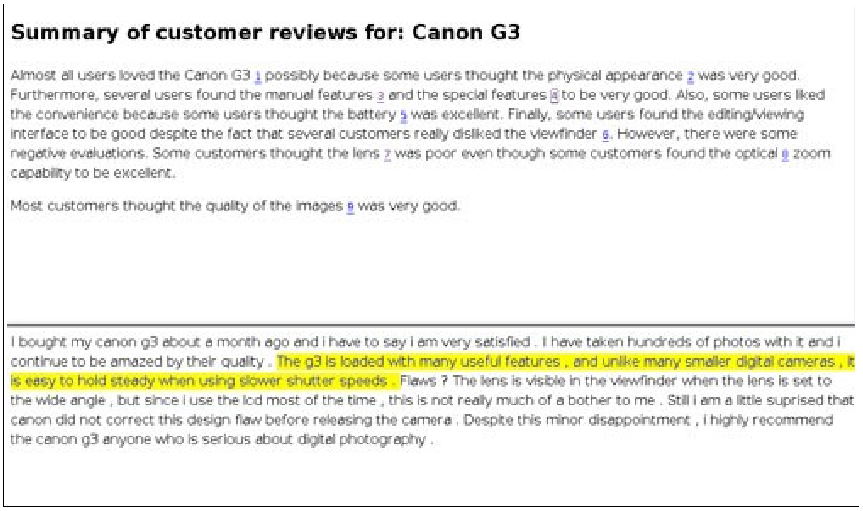
The extractive approach is represented by MEAD*, which is adapted from the open source summarization framework MEAD (Radev et al., 2000).

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 a certain degree of coherence is enforced, as 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 the 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.

2.3.3. Links to the Corpus

In common with the previous study on which this is based (Carenini et al., 2006), both the SEA and MEAD* summaries contain "clickable footnotes" which are links back into an original user review, with a relevant sentence highlighted (Figure 2.2). 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.



Summary of customer reviews for: Canon G3

Almost all users loved the Canon G3 1 possibly because some users thought the physical appearance 2 was very good. Furthermore, several users found the manual features 3 and the special features 4 to be very good. Also, some users liked the convenience because some users thought the battery 5 was excellent. Finally, some users found the editing/viewing interface to be good despite the fact that several customers really disliked the viewfinder 6. However, there were some negative evaluations. Some customers thought the lens 7 was poor even though some customers found the optical 8 zoom capability to be excellent.

Most customers thought the quality of the images 9 was very good.

I bought my canon g3 about a month ago and i have to say i am very satisfied . I have taken hundreds of photos with it and i continue to be amazed by their quality . The g3 is loaded with many useful features , and unlike many smaller digital cameras , it is easy to hold steady when using slower shutter speeds . Flaws ? The lens is visible in the viewfinder when the lens is set to the wide angle , but since i use the lcd most of the time , this is not really much of a bother to me . Still i am a little suprised that canon did not correct this design flaw before releasing the camera . Despite this minor disappointment , i highly recommend the canon g3 anyone who is serious about digital photography .

Figure 2.2: Sample SEA summary showing clickable footnotes linking to a relevant sentence in a user review (highlighted) from the source corpus.

2.4. Measuring Corpus 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.

2.4.1. 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.

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 2.3: 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.*

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.

2.4.2. 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

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

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

Next, we scale this score by the importance of the evaluations divided by the maximum possible importance 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:

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

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.

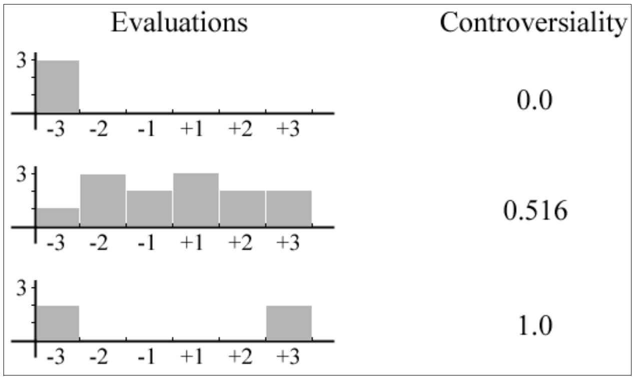


Figure 2.4: Sample feature controversiality scores for three different distributions of polarity/strength evaluations.

2.5. User Study

Our hypothesis is that abstractive summarization outperforms extractive summarization by a larger margin when controversiality is high. We tested this in 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 thirty user reviews per subcorpus was needed to create a summary of sufficient length, which in our case was about 80 words in length.

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 controversiality. Since we would not have been able to test this corpus at high and low levels of controversiality, inclusion of this corpus into the study would have introduced 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, 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 2.3.3), 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. See Appendix A for a copy of the questionnaires.

2.6. Results

2.6.1. 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

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

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 2.1 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 2.2.

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.

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 2.2: Two-way ANOVA p-values.

No significant differences are found in the performance of the summarizers 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 a + in the case where the difference between SEA and MEAD* is greater for a question in the controversial setting, a – if the difference is less, 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 prefer 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 because MEAD* was felt to have made better use of the footnotes than SEA.

2.6.2. 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.”

2.7. Discussion

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. We begin to explore this research problem in the next chapter, in which we examine a framework for content selection that can also select whether content is expressed via extraction or abstraction.

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

3. P-Median for Feature and Strategy Selection

3.1. Clustering As a Framework for Content Selection

We have presented results from a user study in the previous chapter which suggest that the controversiality of opinions in a corpus of evaluative text plays a role in the relative effectiveness of abstractive versus extractive summarization. Specifically, the margin by which abstraction outperforms extraction is greater when controversiality is high. We now present a framework for content selection which is the corpus controversiality by combining abstractive and extractive summarization.

In the summarization systems that were presented and used in the previous section, content selection was based on greedy selection using an importance measure defined for the available features. While this procedure is easy to understand, it has several weaknesses. Greedy selection consists of a series of myopic steps to decide what to include in the summary next, based on what has been selected already and what remains to be selected at this step. Although this series of local decisions may be locally optimal, it may result in a suboptimal choice of contents overall. Another weakness is that a new importance measure must be defined for every summarization system and every summarization task for the system ad hoc. Ideally, we would like a more general content selection framework which requires less modification between tasks.

To address these problems, we propose that content selection for summarization can be viewed as a clustering problem. Intuitively, each sentence or clause in the summary can be thought of as being the summary output that best represents the information content contained in its cluster of information. In this framework, content selection consists of two components. The first is a measure to quantify how well one possible output summary element can express the information content within the source that needs to be expressed, which we call the *information coverage measure*. The second is a clustering paradigm to define clusters of similar information that are expressed by a unit of text in the summary. Once defined, we can leverage extensive existing research on clustering algorithms used in other contexts and applications to solve this clustering problem.

Defining a measure for information coverage is perhaps an easier and less arbitrary task than defining an absolute importance measure of content, because we only need to define the relative semantic distance between the expression and the content expressed. For example, we can rely on similarity metrics such as ones based on distances in WordNet (Fellbaum, 1998). In our work, we use a domain specific hierarchical UDF tree as a guide to define this semantic distance.

Another advantage of this framework is its generality. Because the information coverage measure is between a possible summary output and the information in the source text, we are not limited to selecting information content with the clustering algorithm. We can also select other characteristics of the summary simultaneously, such as the summarization strategy (extractive or abstractive) with which the information content is expressed. Note that we do not necessarily require the final surface realization before content selection, if the information coverage measure does not require it. Rather, we just need the type of output sentence that will surface in the summary.

We will now examine the choices we made for the clustering paradigm and the information coverage measure for our domain of summarizing evaluative text, being sensitive to corpus controversiality.

3.2. P-Median for Clustering

3.2.1. Definition of P-Median

The clustering paradigm that we choose to apply to our problem is the p-median problem (also known as the k-median problem), from facility location theory. In its original interpretation, p-median is used to find optimal locations for opening facilities which provide services to customers, such that the cost of serving all of the customers with these facilities is minimized (Resende and Werneck, 2003). Formally, given:

a set F of m potential locations for facilities,

a set U of n customers,

a function $d : F \; x \; U \rightarrow \Re$ representing the cost of serving a customer with a facility, and

a constant $p \leq m$,

an optimal solution to the p-median problem is a subset S of F, such that

$\sum_{u \in U} \min_{f \in S} d(f, u)$ is minimized, and

$|S| = p$

We reinterpret the p-median problem for the case of content selection for summarization. In the SEA summarizer, the basic unit of content is a clause, which corresponds to one feature of the product being evaluated. Opening a facility in the p-median problem corresponds to selecting a feature to be included in the summary. We later extend this to also include whether the feature is expressed using an extractive or abstractive sentence. Thus, the set F corresponds to all of the features in the product.

The set U consists of the customers we have to serve. In summarization, our goal is to cover the information content contained in the source text, and so serving customers should correspond to covering information. Again, the information in our corpus of user reviews is organized into features, so the set U is the set of product features. The cost function d of serving a customer with a facility is interpreted as the cost of covering a certain feature with some clause, or the *information coverage measure* mentioned above. The better that the clause covers the information contained in a feature, the lower the cost. The constant p is a parameter to the p-median problem, determining how many clauses we wish to select for the final output summary.

3.2.2. Related Problems

There exist related problems in location theory to the p-median problem. The p-centre (or k-centre) problem aims to minimize the maximum distance between the selected location and the customer closest to it (see (Hochbaum and Shmoys, 1986)), or the expression

$$\max_{u \in U} \min_{f \in S} d(f, u)$$

P-centre is not as appropriate for modelling content selection, because we want the minimized cost to take into consideration all of the information content contained in the source text, rather than just the feature that is furthest away semantically from the expression that covers it. The sum in the p-median problem is able to take this into account.

The p-cluster (or k-cluster) problem, by contrast, minimizes the maximum distance between two elements of a cluster, requiring a distance function between the customers rather than between the potential facility locations and the customers. It is also inappropriate, because our notion of cost is the loss of information from the source text in the summary, rather than the diversity of information covered by one expression in the summary.

3.2.3. Computational Complexity of P-Median

Solving the p-median problem is NP-hard in general (Kariv and Hakimi, 1979). If distance function is symmetric, it is solvable in O(pn^2) time for trees, and O(pn) time for paths (Tamir, 1996), where n is the number of locations. In the general symmetric case, the best current theoretical approximation is a 4-approximable algorithm (Charikar and Guha, 1999). For complexity results in the asymmetric case, see (Archer, 2000).

Instead of implementing our own software to solve p-median problems, we rely on an existing implementation, POPSTAR, which uses a hybrid multi-start method (Resende and Werneck, 2004). We use POPSTAR's pmm input format, which can handle arbitrary distance functions. We find that it seems to return the optimal solution for problems of our size in practice.

3.3. Defining Information Coverage

We now present several proposals for reducing the content selection problem to a p-median problem which differ primarily in terms of the grouping of information in the source text and the definition of information coverage. In the first section, we describe a proposal which selects only features, and is a direct alternative to the greedy algorithm used in SEA. In the second and third sections, we extend this by also selecting the strategy (extractive or abstractive) with which the clause describing the selected feature is expressed. This extension, however, is still in its preliminary stage, because of problems of grounding the parameters of the problem to aspects of the summarization process. However, given some reasonable parameters, the algorithm appears to make appropriate choices for trading off extraction and abstraction, indicating that this direction of research is worth pursuing further.

3.3.1. P-Median for Feature Selection

To reduce feature selection to a p-median problem, we need to specify the sets *U*, *F*, and the information coverage measure *d* in terms of properties of the summarization process. For feature selection, the sets *U* and *F* are both simply the set of User Defined Features (UDF) of the product, because the information we need to cover is organized into UDFs, and each output clause describes a UDF.

To specify how well a clause about one feature covers information about another feature, we need to consider both the total amount of information about the covered feature as well as the semantic relationship between the two features. We use the importance measure from section 2.4.2 based on the number and strength of evaluations of the covered feature to quantify the former, but groupin evaluations by UDFs rather than CFs.

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

The UDF tree hierarchy provides a domain-specific mechanism to model the semantic distance between the features. We hypothesize that it is easier for a clause about a more general feature to cover information about a more specific feature than the reverse, and that features that are not in a direct ancestor-descendant relationship cannot cover information about each other because of the tenuous semantic connection between them. Based on these assumptions, we define a multiplier for the above measure of importance based on the UDF tree structure.

Let the potentially selected feature be u_i, the feature to be covered be u_j, and the length of the path in the UDF tree between them be *k*. Then, the multiplier $T(u_i, u_j)$ is defined as follows.

$$T(u_i, u_j) = \left\{ \begin{array}{l} T_{up} \times k\,,\ \ \textit{if } u_i \textit{ is a descendant of } u_j \\ T_{down} \times k\,,\ \ \textit{if } u_i \textit{ is an ancestor of } u_j \\ \infty\,,\ \ \textit{otherwise} \end{array} \right\}$$

T_{up} and *T_{down}* are parameters specifying how difficult it is to cover information in a feature that is an ancestor or descendant of the covering feature. In the example to follow, we pick the values T_{up} = 3 and T_{down}=1, meaning that covering information in an ancestor node is three times more difficult than covering information in a descendant node. For implementation purposes, we use an arbitrarily large positive constant in place of infinity.

A third component of the goodness of coverage specific to the evaluative domain is the distribution of evaluations of the features. Coverage is expected to be less if the distributions are different; for example, if users rated a camera well overall but the zoom poorly, a sentence about how well the camera in general is rated does not provide much evidence that the zoom is not well liked. We use a modified version of the feature controversiality score defined earlier. Given polarity/strength ratings between -3 and +3, we first aggregate the positive and negative evaluations by summing the absolute values of the strengths.

Define:

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

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

Now, calculate the parameter to the Bernoulli distribution corresponding to ratio the importance of the two polarities. That is, Bernoulli with parameter

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

The distribution-based multiplier $D(u_i,u_j)$ is the relative entropy (or Kullback-Leibler divergence) from Ber(θ_i) to Ber(θ_j), plus one for multiplicative identity when the relative entropy is zero.

$$D(u_i, u_j) = \theta_i \log \frac{\theta_i}{\theta_j} + (1-\theta_i) \log \frac{(1-\theta_i)}{(1-\theta_j)} + 1$$

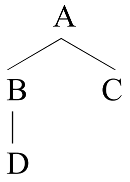
The final formula for the information coverage measure is thus

$$d(u_i, u_j) = imp(u_j) \times T(u_i, u_j) \times D(u_i, u_j)$$

Note that the information coverage measure is not symmetric, because neither relative entropy nor the multipliers are symmetric. Thus, the measure is not a metric or a distance function. This does not pose problems for the POPSTAR implementation.

Example

Consider the following four-node UDF tree and importance scores.



imp(A) = 55

imp(B) = 50 , *imp*(C) = 30

imp(D) = 120

With parameters T_{up} = 3 and T_{down}=1 and setting D to 1, this trees yields the following information measure coverage scores. In the following table, rows represent the covering feature, while columns represent the covered feature.

	covered				
		A	B	C	D
covering	A	0	50	30	240
	B	165	0	∞	120
	C	165	∞	0	∞
	D	330	150	∞	0

Running p-median on these values produces the following optimal results.

p	*Selected*	*Value*
1	A	320
2	A,D	80
3	A,B,D	30
4	A,B,C,D	0

This method trades off selecting centrally located nodes near the root of the UDF tree and the importance of the individual nodes. In this example, D is selected after the root node A even though D has a greater importance value.

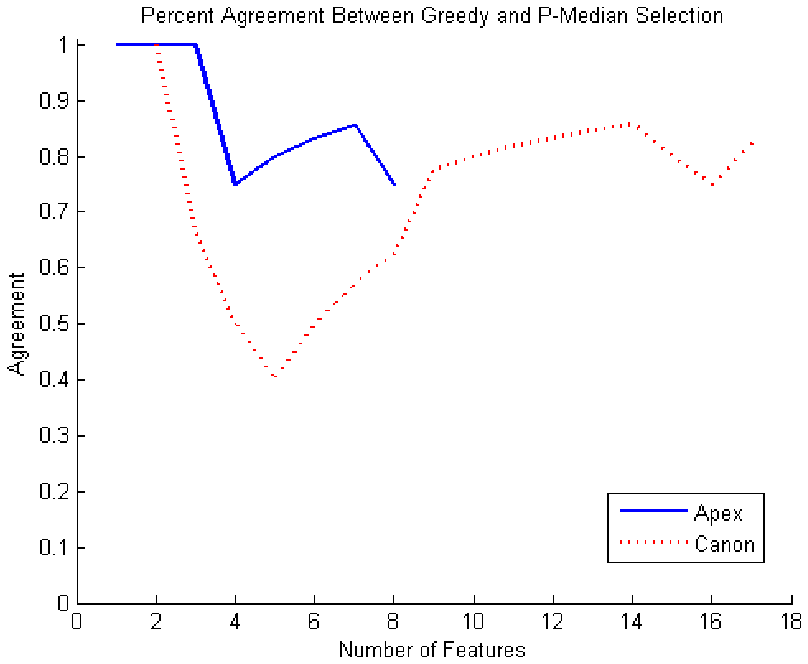


Figure 3.1: Percent agreement between greedy feature selection as found in SEA and p-median based selection.

Agreement with Greedy Selection

As a preliminary form of evaluation, we apply this form of p-median clustering for feature selection, and compare the features selected using this method to the features selected using the greedy method in SEA. Because the user study described in section 2.5 show that content selection was somewhat well received (average content ratings were 3.37 and 3.03 out of 5), we expect a somewhat high correlation between the features selected if p-median is a reasonable form of content selection.

Figure 3.1 shows this comparison on two test corpora for the Apex AD2600 DVD player and the Canon G3 digital camera, from amazon.com. We ran the two selection algorithms selecting different numbers of features, from one feature up to SEA's threshold beyond which SEA would not select any more features. Beyond this point, the remaining features have importance values that are too low to be significant. For the Apex corpus, this threshold was eight features, whereas for the Canon corpus, it was seventeen.

We see that percent agreement is quite high for the Apex corpus, whereas it is low when the number of features is small for the Canon corpus. The agreement quickly rises again after six features, reflecting that the two algorithms select a similar set of features but in a different order. The fact that correlation is not perfect leaves open the possibility that p-median selection outperforms greedy selection. Both curves dip at around four features. It would be interesting to test further, if this dip and then recovery in agreement manifests in many corpora, if they always occur at roughly the same point of the graph, and whether the shape of the curves would be affected by changes to the parameters of the p-median reduction.

3.3.2. Simultaneous Feature and Strategy Selection

The results of the user study described in section 2.5 suggest that abstractive and extractive summarization have different advantages, and that abstractive summarization is preferred when the corpus is controversial. It might thus be advantageous to combine abstractive and extractive elements within a single summary to combine the benefits of each. One problem then is how to determine the proportion of extraction versus abstraction in the summary. One could a priori determine the level of abstraction based on corpus controversiality. The next step would be to specify what information content is expressed with which strategy.

One possibility is to select the features first, and then select a strategy for each feature. However, feature and strategy selection are interdependent. For example, a feature that is especially important might need to be expressed with both an abstractive and an extractive element. If feature and strategy selection were separate phases, it would be difficult to control the number of clauses in the final summary. Another example is that the strategy choice of one feature might affect which of the other features should be chosen.

By using a modified version of the reduction to a p-median problem from the previous section, we can perform feature and strategy selection simultaneously. One drawback is that it is difficult to provide values to many of the parameters in the problem. In the description below, we suggest factors that should influence the choice of values.

In the version that only selects features, the sets F and U correspond to features of the product. Because we now select for both feature and strategy, opening a facility at a location corresponds to selecting a feature to include in the summary as well as the strategy with which to realize it. So, each element of *F* is a feature-strategy pair.

Elements of *U* must also correspond to feature-strategy pairs. In the p-median framework, customers are served by the nearest facility as defined by the facility that can serve them with the lowest cost. It is not possible for a customer to be better served if there are two facilities nearby. In the context of content selection, suppose that a feature is covered by an extractive sentence with cost *dE* and an abstractive sentence with cost *dA*. If both of these sentences were selected, the feature would be considered to be covered with cost min(*dA*,*dE*). It is not possible for the feature to be considered covered with a cost lower than min(*dA*,*dE*), which intuitively is desirable, as it should be possible for two sentences to describe a feature better than one. If elements of *U* correspond to feature-strategy pairs, however, this problem can be solved. Each UDF feature now corresponds to multiple customers in the p-median problem, so if each sentence about a feature reduces the cost of serving one of the customers, the total cost of serving the feature can be further reduced by selecting both strategies.

The next step is to determine how many customers to split each feature into, and how to apportion the information content of the feature among its customers. We divide each feature into three customers, corresponding to three kinds of information that we need to represent about each feature, as well as the strategy that best realizes this information.

(i) *information about the distribution of evaluations* of that feature, best realized by an abstractive clause.

(ii) *information about the details from the positive reviews*, best realized by a positive extractive sentence.

(iii) *information about the details from the negative reviews*, best realized by a negative extractive sentence.

The distinction between (ii) and (iii) is necessary, because an extractive sentence from the corpus is either positive or negative with respect to one feature, and thus only covers information related to the same polarity.

We now define the information coverage measure between feature-strategy pairs. As before, our measure has the following components: a measure of importance based on the number of distribution of evaluations, a multiplier based on the semantic relatedness between features, and a multiplier based on the difference in the distributions of evaluations between the two features. We must also include another multiplier to take into account the effectiveness of one particular strategy covering a certain kind of information.

The measure of importance for a feature is apportioned between the three customers defined above. Factors that might affect this split are feature controversiality (see section 2.4.2), though UDF-based instead of CF-based, the quality of the extractive sentence, and the proportion of positive to negative evaluations. Let $u_i^A, u_i^{E+}, u_i^{E-}$ be the three locations/customers for *udf*$_i$, corresponding to the abstractive, positive extractive, and negative extractive strategies respectively. Then, a reasonable split might be:

$$\begin{aligned}
&imp(u_i^A) = \beta(udf_i) \times imp(udf_i)\\
&imp(u_i^{E+}) = (1-\beta(udf_i)) \times imp\_pos(udf_i)\\
&imp(u_i^{E-}) = (1-\beta(udf_i)) \times imp\_neg(udf_i)
\end{aligned}$$

where *β* is a function based on feature controversiality that splits the importance between the abstractive and extractive locations, and *imp*(*udf*$_i$) is the sum of the polarity/strength evaluations belonging to a UDF node as in section 3.3.1.

The multipliers based on the semantic relatedness of the features and the difference in the distribution of evaluations remain the same. We simply ignore the strategy component of the feature-strategy pair when computing them. The multiplier based on the strategies of the nodes is more difficult to define. A function for determining the multiplier $S(u_i, u_j)$ would require values for the constants in the following table.

$S(u_i, u_j) =$

Strategies for locations i and j	abstractive	positive extractive	negative extractive
abstractive	c_1	c_2	c_3
positive extractive	c_4	c_5	c_6
negative extractive	c_7	c_8	c_9

This table specifies how well a clause of one strategy covers information best covered by another strategy, possibly of another feature. Although we do not have specific values for these parameters, we propose the following heuristics to guide us in defining values for these parameters.

(i) Distribution information is best covered by abstractive sentences, and details by extractive sentences of the same polarity.

(ii) An abstractive clause is better able to cover information outside of its own feature than an extractive clause is able to.

The final information coverage measure is thus

$$d(u_i, u_j) = imp(u_j) \times T(u_i, u_j) \times D(u_i, u_j) \times S(u_i, u_j)$$

Defining the multipliers for tree structure and strategy separately assumes that they are independent. However, it might be the case that the level of asymmetry between covering information in an ancestor versus a descendant node is correlated with the strategies. Furthermore, we assume that an abstractive sentence about the distribution can describe the distribution information better. However, it is possible that an extractive sentence, or multiple extractive sentences in the case of a controversial feature, can do the same.

3.3.3. Extraction as Complementing Abstraction

Instead of attempting to select feature and strategy in general, a more restricted and easier to define possibility is to approach the task of dividing up a feature into multiple customers from the perspective that extraction supports abstraction. That is, we make the assumption that an extractive sentence about a feature can only be present if an abstractive sentence about a feature is also present. This assumption seems reasonable in light of the better performance of SEA in the user study (section 2.6). See Appendix B for a proof that this can be enforced within the p-median framework.

This version of the problem is easier to specify, because instead of asking p-median to select between features and strategies in general, we ask it to optimize between the number of features to express, and the depth to which each feature is expressed using abstraction and possibly extraction. Since extraction is now a more peripheral part of the summary, we must ask what its purposes are in the summary, rather than what kind of information it better represents.

Possible uses for extraction are the following.

(1) *Verification* of information provided by the abstractive sentence as a form of source citation. For example, an extractive sentence can provide evidence to support a claim such as "Some users found the extra features to be very good." in the sample summary in Figure 2.2.

(2) *Elaboration* of information provided by the abstractive sentence, such as what specifically about a feature is liked or disliked by users. In the above example, extraction could provide an instance of a user praising a particular extra feature.

(3) *Complementation* of information not present in the abstractive sentence. Continuing the example, extraction could give a counterexample of a negative evaluation of the extra features.

In the final display of SEA summaries, there are footnotes linking each abstractive clause to a related sentence in the corpus (section 2.3.3). We hypothesize that these footnotes serve the functions of *verification* and *elaboration*. Thus, for an extractive element to be useful in an otherwise abstractive summary, it should primarily serve the function of *complementation*.

We now go through the details of reducing the content selection process to a p-median problem given these assumptions. First, the sets F and U correspond to feature-strategy pairs, but each feature now only has two strategies: extractive and abstractive. The information coverage measure consists of the three components: the measure of importance, the multipliers based on UDF tree structure, and distribution.

The definition of the measure of importance is different. We no longer need to split the overall importance between the two customers, because the two customers are not alternatives for representing the information in a feature. The abstractive importance can be defined as in 3.3.1, whereas the extractive importance is now the importance of the evaluations of orientation *opposite* to the predominant orientation of an uncontroversial node. The threshold of determining whether a UDF is controversial or not is the same as that for determining whether the abstractive clause represents the feature as controversial or uncontroversial.

It is not clear what an appropriate complementation to a controversial feature is. One could provide an extractive element of only one of the two polarities, but as both polarities are in a sense covered by an expression that a feature is controversial, it is not clear that this truly serves the purpose of complementation. In the definitions below, we take the more important of the two polarities to define the importance of an extractive location for a controversial node, in order to preserve having an extractive location for each feature.

Let u_i^A, u_i^E be the abstractive and extractive locations for *udf*$_i$. Then,

$$imp(u_i^A) = imp(udf_i) = \sum_{ps_k \in ps(udf_i)} |ps_k|$$

$$imp(u_i^E) = \left\{ \begin{array}{l} imp\_neg(udf_i), \text{ if } udf_i \text{ predominantly positive} \\ imp\_pos(udf_i), \text{ if } udf_i \text{ predominantly negative} \\ imp\_pos(udf_i), \text{ if } udf_i \text{ controversial, } imp\_pos(udf_i) > imp\_neg(udf_i) \\ imp\_neg(udf_i), \text{ if } udf_i \text{ controversial, } imp\_neg(udf_i) > imp\_pos(udf_i) \end{array} \right\}$$

The multipliers *T(u$_i$, u$_j$)* and *D(u$_i$, u$_j$)* are as described in 3.3.1.

In addition, we must ensure that extractive locations cannot serve any customer other than itself in order to enforce that abstraction is selected before extraction (Appendix B), so we set the corresponding values of the information coverage measure to infinity. The information coverage measure is thus

$$d(u_i, u_j) = \left\{ \begin{array}{l} 0, \text{ if } u_i = u_j \\ \infty, \text{ if } u_i \text{ extractive, } u_i \text{ and } u_j \text{ belong to different features} \\ imp(u_j), \text{ if } u_i \text{ and } u_j \text{ belong to the same feature} \\ imp(u_j) \times T(u_i, u_j) \times D(u_i, u_j), \text{ otherwise} \end{array} \right\}$$

Example

We continue the example from 3.3.1, providing in addition arbitrary though plausible importance values for the extractive locations.

imp	Abstractive (X^A)	Extractive (X^E)
A	55	30
B	50	15
C	30	10
D	120	60

The matrix of information coverage scores is as follows (rows represent the covering node, columns represent the covered node).

		covered							
		A^A	A^E	B^A	B^E	C^A	C^E	D^A	D^E
covering	A^A	0	30	50	15	30	10	240	120
	A^E	55	0	∞	∞	∞	∞	∞	∞
	B^A	165	90	0	15	∞	∞	120	60
	B^E	∞	∞	50	0	∞	∞	∞	∞
	C^A	165	90	∞	∞	0	10	∞	∞
	C^E	∞	∞	∞	∞	30	0	∞	∞
	D^A	330	180	150	45	∞	∞	0	60
	D^E	∞	∞	∞	∞	∞	∞	120	0

Below are some sample calculations illustrating the calculation of *d*.

Case 1 – *d*(B^E,D^A)

d(B^E,D^A) = ∞, as extractive locations cannot cover locations belonging to other features.

Case 2 – *d*(A^A,D^E)

$$d(A^A, D^E) = imp(D^E) \times T(A^A, D^E) \times D(A^A, D^E)$$
$$d(A^A, D^E) = 60 \times (T_{down} + T_{down}) \times 1 = 60 \times 2 = 120$$

Case 3 – *d*(D^A,A^A)

$$d(D^A, A^A) = imp(A^A) \times T(D^A, A^A) \times D(D^A, A^A)$$
$$d(D^A, A^A) = 55 \times (T_{up} + T_{up}) \times 1 = 55 \times 6 = 330$$

Case 4 – *d*(B^A,C^E)

d(B^A,C^E) = ∞, because B and C are sibling nodes, and are therefore not ancestors or descendants of each other.

Running p-median on these values produces the following optimal results.

p	*Selected*	*Value*
1	A^A	495
2	A^A, D^A	195
3	A^A, D^A, D^E	135
4	A^A ,B^A ,D^A ,D^E	85
5	A^A, A^E, B^A, D^A, D^E	55
6	A^A, A^E, B^A, C^A, D^A, D^E	25
7	A^A, A^E, B^A, B^E, C^A, D^A, D^E	10
8	A^A, A^E, B^A, B^E, C^A, C^E, D^A, D^E	0

We see that the abstractive locations are selected in the same order as before, but extractive locations are also selected, sometimes before abstractive locations of different features, as is the case with D^E, selected before B^A. In this example, all nodes selected with a lower *p* are also selected when *p* is greater. However, this is not necessarily the case in general.

3.4. Discussion and Future Work

In this section, we proposed a novel framework for summarization content selection based on clustering. In this framework, content selection is viewed as selecting clusters of information, where each cluster consists of related information that is represented by one unit of text in the output summary. We applied the framework to the specific context of the evaluative summarization of customer reviews using the p-median clustering problem.

Mapping content selection into a p-median problem gives us the flexibility to select other properties of the summary such as the strategy along with the features simultaneously. More investigation is needed to define the potential facility locations and customers, as well as the information coverage between them in a natural manner grounded in properties of the corpus.

The p-median framework also imposes limitations, one of which is that the each facility, or output unit of text, takes up exactly one of the *p* slots. Because we cannot specify variable costs for opening facilities, we cannot, for example, bundle two extractive sentences to be selected together, which take up two of the *p* slots for facilities.

Our approach currently lacks a real evaluation of its effectiveness as compared to the greedy methods presently being used in the summarizers. There are several possibilities for evaluating them. One is to directly evaluate the features that the two algorithms select, by running a user study to determine which features users would prefer to include in a summary. Another is to generate summaries based on these methods and asking users to evaluate the summaries.

In the approaches proposed in sections 3.3.2 and 3.3.3, we allow a feature to be expressed by an extractive or abstractive element. This could lead to coherency issues, if, for example, an extractive sentence is embedded between abstractive sentences without context. We will likely need to introduce extractive elements in some way. One possibility is to treat them as direct quotations with introductory clauses. e.g. “One customer said that, 'The player broke two months after I bought it.'” Another possibility is to introduce them as indirect quotations. e.g. “One customer said that the player broke two months after he bought it.” This will require more linguistic modification of the extractive elements, such as rewriting first to third person, resolving anaphoric references, and eliminating references to a portion of the user review outside of the quoted text. Although these modifications mean that the extractive elements are not purely extractive, they preserve for the most part the characteristics of extraction, because they are closely based on original source text. We thus expect them to behave similarly to purely extractive elements for the purposes of information coverage, linguistic fluency, etc.

4. Conclusion

In this thesis, we examined two issues related to the summarization of evaluative text containing opinions and preferences. First, we defined and investigated the effect of the controversiality of opinions in the corpus on two strategies of summarization—abstractive summarization with novel text and extractive summarization where a summary is composed of extracts from the source corpus. The results of a user study we ran suggest that abstraction outperforms extraction by a greater margin when the controversiality of the corpus is high. Based on this result, we hypothesized that a more effective approach to summarization would combine extraction and abstraction to leverage the advantages of each, and that controversiality would be a factor determining the mix of abstraction and extraction in the summary.

We then proposed a summarization content selection framework based on clustering, and applied this framework to our evaluative domain. In this framework, clusters represent groupings of related information that are represented or covered by a unit of text in the output summary. Using the p-median problem from facility location theory as the clustering paradigm, we detailed several methods of reducing the content selection problem to a p-median problem. These methods can select not only the features to be realized in the summary, but also the strategy with which to express them.

More work is needed to specify the parameters to the resultant p-median problems, especially in the more general versions of the reduction which allow any combination of features and strategies to express the content. For example, in one version of the reduction, a measure of information associated with each feature must be divided into buckets according to the summarization strategy that best represents that information. What specific properties in the corpus should determine how this split is performed remains to be investigated.

There are also unresolved issues associated with combining extraction and abstraction after the content selection stage. One is the effect that this combination might have on the coherency of the summary, and how extractive elements should be realized to maintain coherency. Another issue is that the hybrid summaries must still be evaluated to determine whether there is any improvement in performance over purely extractive or purely abstractive summaries.

References

Aaron Archer. 2000. Inapproximability of the asymmetric facility location and k-median problems. Unpublished manuscript.

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

Giuseppe Carenini and Jackie C. K. Cheung. To appear, 2008. Extractive vs. NLG-based Abstractive Summarization of Evaluative Text: The Effect of Corpus Controversiality. *Fifth INLG 08*, 8 pages, Salt Fork, OH.

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

Giuseppe Carenini, Raymond T. 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.

M. Charikar and S. Guha. 1999. Improved combinatorial algorithms for the facility location and k-median problems. In *Proc. 40th Annual IEEE Symposium on Foundations of Computer Science*, pages 378–388.

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*, 120-127.

Christiane Fellbaum. (ed.) 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Massachusetts.

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*.

Dorit S. Hochbaum and David B. Shmoys. 1986. A unified approach to approximation algorithms for bottleneck problems. Journal of the ACM 33, 533–550.

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

O. Kariv and S.L. Hakimi. 1979. An algorithmic approach to network location problems, Part II:

p-medians. *SIAM Journal of Applied Mathematics*. 37, 539-560.

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*.

Mauricio G. C. Resende and Renato F. Werneck. 2004. A hybrid heuristic for the p-median problem. *Journal of Heuristics*, 10(1), 59–88.

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

Arie Tamir. 1996. An O(pn^2) algorithm for the p-median and related problems on tree graphs, *Operations Research Letters.* 19, 59–94.

Appendices

Appendix A. 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.

Appendix B. Enforcing Abstraction Before Extraction

We can enforce that the abstractive node of a feature is always selected before the extractive node of the same feature, given several constraints and assuming that the p-median problem is solved optimally.

Let *A* and *E* be the abstractive and extractive nodes for a feature, having set up the p-median problem as described in section 3.3.3.

Constraints

(i) The cost of serving a node from itself is 0.

(ii) The cost from E to any node other than itself is greater than or equal to the cost from A to the same node.

(iii) The cost from A to E is smaller than the cost from E to A, and these costs are the smallest, other than from the node to itself.

Theorem

E will never be selected without A also being selected.

Proof by Contradiction

Suppose E is selected without A in an optimal solution with the value of v_{old}, where d^{X}_{old} is the cost of serving node X. *OTHERS* refers to the set of customer nodes other than nodes A or E.

$v_{old} = d^{OTHERS}_{old} + d^{A}_{old} + d^{E}_{old}$, by definition

$= d^{OTHERS}_{old} + d^{A}_{old}$, by constraint (i)

Substitute A for E in the selected set of nodes, with the value of v_{new}, where d^{X}_{new} is the cost of serving node X.

$v_{new} = d^{OTHERS}_{new} + d^{A}_{new} + d^{E}_{new}$, by definition

$= d^{OTHERS}_{new} + d^{E}_{new}$, by constraint (i)

$d^{OTHERS}_{new} \leq d^{OTHERS}_{old}$, otherwise (ii) would be violated.

$d^{E}_{new} < d^{A}_{old}$, otherwise (iii) would be violated.

then $v_{new} < v_{old}$, which contradicts optimality. ■