Monotone Submodularity in Opinion Summaries

Jayanth Jayanth
IIT Bombay
jayanthjaiswal10
@cse.iitb.ac.in

Jayaprakash Sundararaj
IIT Bombay
osjayaprakash
@gmail.com

Pushpak Bhattacharyya
IIT Bombay
pb
@cse.iitb.ac.in

Abstract

Opinion summarization is the task of producing the summary of a text, such that the summary also preserves the sentiment of the text. Opinion Summarization is thus a trade-off between summarization and sentiment analysis. The demand of compression may drop sentiment bearing sentences, and the demand of sentiment detection may bring in redundant sentences. We harness the power of submodularity to strike a balance between two conflicting requirements. We investigate an incipient class of submodular functions for the problem, and a partial enumeration based greedy algorithm that has performance guarantee of 63%. Our functions generate summaries such that there is good correlation between document sentiment and summary sentiment along with good ROUGE score, which outperforms the-state-of-the-art algorithms.

1 Introduction

Sentiment Analysis is often addressed as a classification task, which aims at determining the sentiment of a word, sentence, paragraph or a document as a whole into positive, negative or neutral classes (Pang et al., 2002). Summarization, on the other hand is the task of aggregating and representing information content from a single document or multiple documents in a brief and fluent manner. Due to the explosive growth of data, fine grained sentiment analysis as well as summarization on the whole chunk of data can be a very time-consuming task. Sentiment Analysis also requires filtering of text portions as either objective (factual information) or subjective (expressing some sentiment or opinion) during pre-processing and then, classifying the subjective extracts as positive or negative. Subjective extracts can also be provided to users as a summary of the sentiment-oriented content of the reviews in search engines. In this paper, we address the problem of generic extractive summarization of reviews, a task commonly known as Opinion Summarization (Liu, 2012). The goals of opinion summarization are:

1. Present a short summary that conveys the essence as well as the sentiment of the review
2. Provide a short subjective extract to NLP pipeline for faster execution (*e.g.* sentiment analysis, review clustering *etc.*).

In this paper, we use movie reviews for opinion summarization task as they often have the following parts:

1. Plot - Description of the story, which is factual in nature
2. Critique - Opinion about the movie, which is sentiment bearing

Clearly, opinion summary to be generated will have a trade-off between the two opposing parts - subjective critique and objective plot. Our goal is to strike a balance through linear combination of suitable submodular functions in our paper. Joint models of relevance and subjectivity have a great benefit in that they have a large degree of freedom as far as controlling redundancy goes. In contrast, conventional two-stage approach Pang and Lee (2004), which first generate candidate subjective sentences using min-cut and then selects top subjective sentences within budget to generate a summary, have less computational complexity than joint models. However, two-stage approaches are suboptimal for text summarization. For example, when we select subjective sentences first, the sentiment as well information content may become redundant for a particular aspect. On the

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 169–178,
Lisbon, Portugal, 17-21 September 2015.

other hand, when we extract sentences first, an important subjective sentence may fail to be selected, simply because it is long. The two stage conflict in the sense that the demand of compression may drop sentiment bearing sentences, and the demand of sentiment detection may bring in redundant sentences. We then, use partial enumeration based greedy algorithm (Khuller et al., 1999), which gives performance guarantee of $(1 - e^{-1}) \approx 0.632$ (Sviridenko, 2004). The performance guarantee reported is better than simple greedy algorithm, used by Lin and Bilmes (2010) as their proof is erroneous (Morita et al., 2013). Further, the same greedy algorithm, which was used again in Lin and Bilmes (2011) gives only performance guarantee of $\frac{1}{2}(1 - e^{1}) \approx 0.316$ (Khuller et al., 1999).

The rest of the paper is as follows - in the next section, we look at previous work and establish further motivation for our work. Following that, we build the theory and formulate suitable objectives for opinion summarization task. In the final section, we present results based on implementation and testing of the functions. Experimental results show that the functions outperform the-state-of-the-art methods.

2 Previous Work

Automatically generating opinion summaries from large review text corpora has long been studied in both information retrieval and natural language processing.

In (Pang and Lee, 2004), a mincut-based algorithm was proposed to classify each sentence as being subjective or objective. The purpose of this work was to remove objective sentences from reviews to improve document level sentiment classification. Interestingly, the cut functions are symmetrical and submodular, and the problem of finding min-cut is equivalent to minimizing a symmetric submodular function.

Lerman et al. (2009) proposed three different models - sentiment match (SM), sentiment match + aspect coverage (SMAC) and sentiment-aspect match (SAM) to perform summarization of reviews of a product. The first model is called sentiment match (SM), which extracts sentences so that the average sentiment of the summary is as close as possible to the average sentiment rating of reviews of the entity *i.e.* low MISMATCH but with high sentiment INTENSITY. The second model, called sentiment match + aspect coverage (SMAC), builds a summary that trades-off between DIVERSITY, maximally covering important aspects and MISMATCH, matching the overall sentiment of the entity along with high INTENSITY. The third model, called sentiment-aspect match (SAM), not only attempts to cover important aspects, but cover them with appropriate sentiment using KL-Divergence function. Here, INTENSITY and DIVERSITY in the first two models are linear monotone submodular functions, while KL-Divergence function *i.e.* relative entropy in last model, unlike entropy is not monotone submodular.

In (Nishikawa et al., 2010b), a more sophisticated summarization technique was proposed, which generates a traditional text summary by selecting and ordering sentences taken from multiple reviews, considering both informativeness and readability of the final summary. The readability score in this paper would have been linear monotone submodular function, if the negative polarity was not penalizing. In (Nishikawa et al., 2010a), the authors further studied this problem using an integer linear programming formulation.

On the other hand, Lin et al .(2011) treated the task of generic summarization as monotone submodular function maximization. Further, they argued that monotone non-decreasing submodular functions are an ideal class of functions to investigate for document summarization. They also show, in fact, that many well-established methods for summarization (Carbonell and Goldstein, 1998; Filatova, 2004; Riedhammer et al., 2010) correspond to submodular function optimization, a property not explicitly mentioned in these publications. Since many authors either in summarization or opinion summarization have used functions similar to submodular functions as objective, we can take this fact as testament to the value of submodular functions for opinion summarization.

3 Theoretical Background

3.1 Introduction to Submodular Functions

A submodular function is a set function ($f : 2^V \rightarrow R$) having a natural diminishing returns property. Diminishing returns property holds if the difference in the value of the function that a single element makes when added to an input set decreases as the size of the input set increases *i.e.* for every $A, B \subseteq V$ with $A \subseteq B$ and every $x \in V \backslash B$, we have that $f(A \cup \{x\}) - f(A) \geq f(B \cup \{x\}) - f(B)$.

A submodular function f is monotone if for every $A \subseteq B$, we have that $f(A) \leq f(B)$.

The extractive summarization task can be modeled as optimization problem *i.e.* finding a set $S \subseteq V$ (S is set of sentences in summary, V is set of sentences in Document) which maximizes a submodular function $f(S)$ subject to budget constraints. In the following section, we will justify the use of submodular function for opinion summarization. Another advantage of choosing monotone submodular function is that there exists a polynomial-time greedy algorithm for constrained monotone submodular objective. The greedy algorithm guarantees that the summary solution obtained is almost as good as (63%) the best possible summary solution according to the objective (Sviridenko, 2004; Wolsey, 1982).

3.2 Submodularity in Opinion Summarization

Opinion Summarization should be modeled as a monotone submodular optimization problem, since opinion summary also holds following properties:

1. Monotonicity - As more sentences are added to opinion summary, subjectivity increases along with information content as opinionated words are being added.

2. Diminishing Return - If multiple sentences of varying intensity are added to opinion summary, the effect of a lower intensity polarity bearing sentence is diluted in the presence of a higher intensity one.

To show that opinion summarization inherently follow the diminishing return property, consider the following sentences[1] with positive polarity:
A: "*Even the acting in From Hell is solid, with the dreamy Depp turning in a typically strong performance and deftly handling a British accent.*"
B: "*Worth mentioning are the supporting roles by Ians Holm and Richardsonlog.*"
(A ∪ B) : "*Even the acting in From Hell is solid, with the dreamy Depp turning in a typically strong performance and deftly handling a British accent. Worth mentioning are the supporting roles by Ians Holm and Richardsonlog.*" Compare A and its superset, A ∪ B as candidate summaries. Sentence A and B convey positive sentiment, but sentence B has less intensity compared to sentence A. After reading the text (A ∪ B), it is clear that the effect of sentence B has diminished in front of sentence A, though both are of same polarity. B can be thus, removed from the candidate summary as it does a diminishing addition in presence of sentence A to the positive sentiment over the "acting" aspect of the entity "movie". The diminishing return not only holds for same polarity but also, for opposite polarity. Consider another example[2]:
A: "*The movie is predictive with foreseeable ending.*"
B: "*Still it's very well-done that no movie in this entire year has a scene that evokes pure joy as this does.*"
(A ∪ B) : "*The movie is predictive with foreseeable ending. Still it's very well-done that no movie in this entire year has a scene that evokes pure joy as this does.*" Compare B and its superset, A ∪ B as candidate summaries. Sentence A has negative sentiment whereas sentence B conveys positive sentiment with more intensity. When we read the text (A ∪ B), it is clear that the effect of sentence A has diminished in front of sentence B in text , as usually polarity of higher intensity dominates over the polarity of lower intensity. Now, consider a general example[3],
"*Laurence plays Neo's mentor Morpheus and he does an excellent job of it. His lines flow with confidence and style that makes his acting unique and interesting. The movie has lot of special effects and action-packed scenes with part of the appeal has philosophical and religious underpinnings.*"

If the budget for summary had been only two subjective sentences, then picking up first two would have redundantly captured only single aspect (*i.e.* acting) and the redundancy of the concept (acting) also causes a diminishing return of the second sentence because of the difference in sentiment intensity. However, picking the last sentence with either one of the first two would have not just covered both the aspects (*i.e.* acting and visual effects) but since, the sentences are not overlapping in aspects, there would not have been any diminishing return of sentiment on shared aspect (acting). Thus, it can be verified that opinion polarity also holds submodular property of diminishing return, if they are on the same aspect of a distinct entity.

[1] http://www.imdb.com/reviews/295/29590.html

[2] http://www.imdb.com/reviews/159/15918.html

[3] http://www.imdb.com/title/tt0133093/reviews

4 Formulation

Let V represent the set of the sentences in a document. The task of extractive opinion summarization is to select a subset $S \in V$ to represent the entirety (ground set V) . Obviously, we should have $|S| \leq |V|$ as it is a summary and should be small. Therefore, constraints on S can naturally be modeled as knapsack constraints:

$$\sum_{i \in S} c_i \leq b \tag{1}$$

where c_i is the non-negative cost of selecting unit i (*e.g.*, the number of words in the sentence) and b is our budget. If we use a set function $F : 2^V \rightarrow R$ to measure the quality of the summary set S, the summarization problem can then be formalized as the following combinatorial optimization problem:

$$S^* \in argmax_{S \subset V} F(S) \textit{ s.t. } \sum_{i \in S} c_i \leq b \tag{2}$$

where F (S) , total utility of summary is given as a linear combination of L(S), relevance and A(S), subjective coverage of aspects.

$$F(S) = \alpha L(S) + \beta A(S) \tag{3}$$

This formulation clearly brings out the trade-off between the subjective and the objective part. The intuition behind the combination of sentiment and aspect coverage in same function A(S) is that opinion polarity holds submodular property of diminishing return only if the set of sentences talk about common aspect of the same entity as discussed in previous section. L(S) , relevance is modeled same as in (Lin and Bilmes, 2011) as it captures the summary property, while our novel function, A(S) has been modeled differently through a suitable submodular function such that it captures the subjectivity property.

$$L(S) = \sum_{i \in V} min\{c_i(S), \gamma c_i(V)\} \tag{4}$$

$$c_i(S) = \sum_{j \in S} w_{i,j} \tag{5}$$

Here, $w_{i,j} > 0$ measures the similarity between i^{th} and j^{th} sentences and $c_i(S)$ measures the similarity of summary with the document.

Since, A(S), subjective coverage of aspects has to be modeled as monotone submodular function, it has been formulated as :

1. A_1 : **Modular Function**

   A_1(S) is simple linear function, which is sum of weighted subjective scores for each sentence. No budgeting constraints are added to this formulation.

   $$A_1(S) = \sum_i \sum_{j \in (P_i \cap S)} s_j * w_i \tag{6}$$

   Here $P_i; i = 1...K$ is a partition of the ground set V (*i.e.*, $\cup_i P_i = V$), which contains sentences pertaining to different distinct aspects. w_i are the weights of the partitions, based on the corresponding aspects. s_j is the subjective score of the sentence j in summary. The subjective score s_j is calculated using sentiwordnet as sum of the positive score $\in [0, 1]$ and negative score $\in [0, 1]$ (Esuli and Sebastiani, 2006).

   $$s_j = \sum_{word \in j} (pos(word) + neg(word)) \tag{7}$$

2. A_2 : **Budget-additive Function**

   $A_2(S)$ is an extension to $A_1(S)$, where maximum subjectivity score is restricted with budget based on aspect. Here, $\lambda \in [0, 1]$ is threshold coefficient for budget additive function to avoid redundancy of high sentiment on same aspect. When aspect i is saturated by S ($min(\sum_{j \in (P_i \cap S)} s_j, \lambda) = \lambda$), any new sentence j cannot further improve coverage over i and thus, other aspects, which are not yet saturated will have a better chance of being covered. This formulation ensures that produced summary is diverse enough and conveys sentiment about different aspects by budgeting.

   $$A_2(S) = \sum_i min(\sum_{j \in (P_i \cap S)} s_j, \lambda_i) * w_i \tag{8}$$

3. A_3 : **Polarity Partitioned Budget-additive Function**

   In previous formulation we have not considered the polarity of the sentences. For example, if an aspect have many positive sentences with more intensity but few negative

sentences with less intensity, A_2 more likely to reward more positive sentences because of intensity. In this formulation budgeting applied not only on aspect but polarity scores too. This ensures that both positive and negative polarity sentences are part of summary.

$$A_3(S) = \sum_i min(\sum_{j \in (P_i \cap S \cap P_{pos})} s_j, \lambda_i) * w_i + min(\sum_{j \in (P_i \cap S \cap P_{neg})} s_j, \lambda_i) * w_i \quad (9)$$

P_{pos} and P_{neg} are the partition of the sentences in the ground set V, based on their sign of polarity score. The polarity score pol_j for partitioning sentences into P_{pos} and P_{neg} is calculated as difference of the positive and negative score.

$$pol_j = \sum_{word \in j} (pos(word) - neg(word)) \quad (10)$$

Polarity based partitions bring out contrast view on a particular aspect, which is similar to contrast view opinion summarization to give the reader a direct comparative view of different strong opinions.

4. A_4 : **Facility Location Function**

   In this formulation, we model the facility location objective function (Krause and Golovin, 2014) for opinion summarization as choosing possible sentences (facilities) out of document (set of locations) to serve aspects (customers) giving service of value s_j. If each aspect (customer) chooses the sentences (facility) with the highest value, the total value provided to all aspects (customers) is modeled by this set function.

$$A_4(S) = \sum_i max_{j \in (P_i \cap S)} s_j * w_i \quad (11)$$

So A_4 rewards only a sentence which has maximum subjectivity score in each aspect.

5. A_5 : **Polarity Partitioned Facility Location Function**

   A_5 is similar to A_4, but for each aspect, A_5 rewards two sentences with positive and negative polarity but with maximum subjectivity scores in those polarity partitions.

$$A_5(S) = \sum_i max_{j \in (P_i \cap S \cap P_{pos})} s_j * w_i + \sum_i max_{j \in (P_i \cap S \cap P_{neg})} s_j * w_i \quad (12)$$

Each of the above functions are monotone submodular as the parameters s_j and w_i are positive. Since the first function is linear, it is both submodular and supermodular, thus modular. Budget additive and facility location functions (Krause and Golovin, 2014) are special types of monotone submodular functions. Since, monotone submodularity is preserved under non-negative linear combinations, polarity based partitioned function, whose sub-parts are monotone submodular is also monotone submodular.

5 Experiment

We have created Movie ontology tree manually (figure 1). Further the ontology is enriched by adding clue words to all aspects using wordnet sense propagation algorithm (Esuli and Sebastiani, 2006) for three iterations. The algorithm does a hard clustering of the sentences by assigning the sentence aspect, which has maximum number of clue words in that sentence. Clue words for 'Plot' aspect are *story, script, storyline, chief, communicative, explain, narrate, narration, narrative, narrator, report, reporter, scheme, schemer, script, scriptural, storyteller, tell, write up..,*.

For the experiments, we have used the polarity dataset from Pang et al. (2004). The dataset contains 1000 positive and 1000 negative movie reviews with size varying between 700 to 1000 words. As summary generation is time consuming task (DUC[4] only used 25 summaries to evaluate the performance of systems), we picked 100 positive and 100 negative reviews randomly from the dataset and their abstract summaries are generated manually with 200 words limit as budget for

[4]Document Understanding Conferences, http://duc.nist.gov

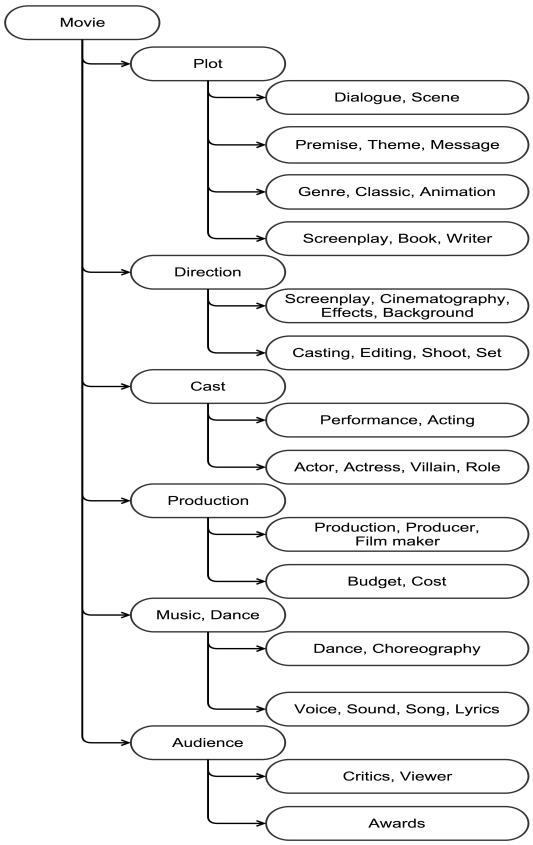

Figure 1: Movie Ontology Tree

evaluation. These 200 summaries are used as gold standard for estimating ROUGE scores of system generated summaries.

In the experiment, the partial enumeration based greedy algorithm (Khuller et al., 1999) is used for summary generation of 200 test documents within budget of 200 words. The algorithm has two parts. In the first part, the algorithm compares function values of all feasible solutions (sets) of cardinality one or two. Let $Summ_1$ be a feasible set of cardinality one or two that has the largest value of the objective function $F(S)$. In the second part, the algorithm enumerates all feasible sets of cardinality three. The algorithm, then completes each such set greedily and keeps the current solution feasible with respect to the knapsack constraint. Let $Summ_2$ be the solution obtained in the second part that has the largest value of objective function over all choices of the starting set for the greedy algorithm. Finally, the algorithm outputs $Summ_1$ if $F(Summ_1) > F(Summ_2)$ else $Summ_2$ otherwise. The algorithm does O(n^2) function calculations in first part, while O(n^5) in second part. This algorithm gives a performance guarantee of $(1-e^{-1})$ for solving monotone submodular objective with knapsack constraint (Khuller et al., 1999; Sviridenko, 2004). As far as we know, the algorithm has not been implemented for such problems because of complexity constraints (Lin and Bilmes, 2011).

Algorithm 1 Overall Algorithm - Summary Extraction

$B \Leftarrow 200$
for *Sentence s* $\in$ *Document* V **do**
  Assign sentence s to one of aspects in movie ontology.
end for
$Summ_1 \Leftarrow$ argmax { F(S), such that S $\subseteq$ V, |S| < 3, and cost(S) $\leq$ B }
$Summ_2 \Leftarrow \emptyset$
for all S $\subseteq$ V, |S|=3, and cost(S) $\leq$ B **do**
  $U \Leftarrow V \backslash S$
  while U $\neq \emptyset$ **do**
    maxReturn $\Leftarrow$ *0.0*
    newSentence $\Leftarrow \emptyset$
    for *Sentence s* $\in$ U **do**
      $S^* \Leftarrow S \cup \{s\}$
      $F(S^*) \Leftarrow \alpha L(S^*) + (1-\alpha)A(S^*)$
      $return \Leftarrow \frac{F(S^*)-F(S)}{len(s)}$
      if $return \geq maxReturn$ **then**
        $maxReturn \Leftarrow return$
        $newSentence \Leftarrow s$
      end if
    end for
    if cost($S \cup \{newSentence\}$) $\leq$ B **then**
      $S \Leftarrow S \cup \{newSentence\}$
    end if
    U $\Leftarrow$ U $\backslash \{newSentence\}$
  end while
  if F(S) $\geq F(Summ_2)$ **then**
    $Summ_2 \Leftarrow$ S
  end if
end for
if F($Summ_1$) $\geq$ F($Summ_2$) **then**
  Summary $\Leftarrow Summ_1$
else
  Summary $\Leftarrow Summ_2$
end if

In the algorithm, the sentences are clustered in different partitions, corresponding to different aspects in the ontology tree using the clue words. In the experiment, hard clustering of the sentences in aspect-based partitions is considered but soft-clustering of the sentences will also work with this approach, which has been left out to avoid further parameter tuning for soft clustering assigments. The weights of the partitions as well as the threshold parameters for the A(S) are currently kept proportional to the inverse of the depth of that aspect in the ontology-tree as sentiment expressed on the concepts at higher level in the ontology tree should have more weightage.

$\forall$ Aspects i,

$$w_i = \lambda_i = \frac{1}{Level(i)}$$

The linear combination parameter β is set as $1-\alpha$ to bring out the trade-off between relevance and subjective coverage of aspects and α is varied from 0 to 1 with step size 0.05 to find optimal α. γ in L(S) is set to 0.5. The parameter learning, esp. α and its impact have been already studied in (Lin and Bilmes, 2011) and thus, is not addressed in the paper. We have used the same approach of grid search to find the optimal value of α.

6 Results

We use ROUGE (Lin, 2004) for evaluating the content of summaries. We have used the 200 test documents that are manually summarized as gold standard data for ROUGE evaluation. For figuring out the sentiment correlation between manual and system generated summaries, we trained Naive Bayes sentiment classifier (Pang et al., 2002) on training data using bag of words approach with features as unigrams and bigrams and then, using minimum Pearson's chi-square score of 3 for feature extraction (Pecina and Schlesinger, 2006) before calculating the sentiment. The measure of sentiment preservation is calculated as Pearson correlation between the sentiment score of the document and the corresponding summary sentiment, both calculated by the Naive Bayes sentiment classifier while the measure of coverage of information content is given by ROUGE-1 and ROUGE-2 f-scores. Mathematically,

$$Correlation(X,Y) = \frac{Covariance(X,Y)}{std.dev(X) * std.dev(Y)} \quad (13)$$

Here, random variable X is the sentiment score of the document sample and random variable Y is the sentiment score of the corresponding summary sample. For 200 documents, it will be $[(X_1,Y_1),(X_2,Y_2),...,(X_{200},Y_{200})]$ sample points for the above correlation function.

Following five baselines are used for comparison:

1. **Baseline-1/TOP** : Sentences selected consecutively from the start of the review within the budget.

2. **Baseline-2/TOP-SUBJ** : Sentences ranked based on their subjectivity and then, selected with in the budget.

3. **Baseline-3/LER-SM** : (Lerman et al., 2009) Sentences which have sentiment close to document sentiment are chosen as Summary. We have used same NaiveBayes classifier (Pang et al., 2002) trained on imdb corpus to predict the sentiment of a sentence and document.

$$min_{S \subset V} \sum_{j \in S} (|senti(V) - senti(j)|) \quad (14)$$

4. **Baseline-4/TEXTRANK** : TextRank summarizer is based on Graph based unsupervised algorithm. Graph is constructed by creating a vertex for each sentence in the document and edges between vertices based on the number of words two sentences (of vertices) have in common and then, ranking them by applying PageRank to the resulting graph. Summary is generated with sentences having more vertex score (Mihalcea and Tarau, 2004).

5. **Baseline-5/MINCUT** : Mincut algorithm (Pang and Lee, 2004) classifies the sentences as subjective and objective sentences, by finding minimum s-t cuts in graph of sentences using maximum flow algorithm. In the graph, each sentence is a vertex and the edge between the vertex to the source or sink is taken as probability of the sentence being subjective or objective (individual scores). To ensure the graph connectivity, edges are drawn between every pair of sentence vertices, with edge weights taken proportional to the degree of proximity (association scores). After maximum flow algorithm, the cut in which source vertex lies is classified as subjective and vice-versa. We pick top subjective sentences within the budget as summary.

Among the five baselines, TOP and TOP-SUBJ are simplistic. Though both TEXTRANK and MINCUT were not originally proposed for opinion summarization but a number of papers in opinion summarization have built over these two methods and also, used them as baselines and thus, comparing with the "well-known" baselines will give the readers from the sentiment analysis field an intuitive idea of the performance of our system. MINCUT, however was reproposed specifically for subjective summarization by Pang and Lee (2004) and we use that formulation for comparison.

Table 1 compares the five functions with the above baselines based on optimal values of tradeoff α. From the table, it can be inferred that all the

System	ROUGE1	ROUGE2	S. Corr.
TOP	0.43001	0.16591	0.86144
TOP-SUBJ	0.41807	0.14362	0.82953
LER-SM	0.42608	0.14533	0.96545
TEXTRANK	0.41987	0.14644	0.88967
MINCUT	0.39368	0.11047	0.84017
Submod-A_1	0.43223	0.15702	0.95306
Submod-A_2	0.43594	0.15977	0.97538
Submod-A_3	0.43247	0.15436	0.93155
Submod-A_4	**0.43602**	**0.15760**	**0.98566**
Submod-A_5	0.42976	0.15551	0.95415

Table 1: ROUGE F-score and sentiment correlation for optimal values of α with baselines 1-5

System	ROUGE1	ROUGE2	S. Corr.
LIN-A_1	0.43112	0.15795	0.89850
LIN-A_2	0.42704	0.15382	0.90212
LIN-A_3	0.42612	0.15297	0.93155
LIN-A_4	0.42688	0.15245	0.93905
LIN-A_5	0.43359	0.15922	0.91019
Submod-A_1	0.43223	0.15702	0.95306
Submod-A_2	0.43594	0.15977	0.97538
Submod-A_3	0.43247	0.15436	0.93155
Submod-A_4	**0.43602**	**0.15760**	**0.98566**
Submod-A_5	0.42976	0.15551	0.95415

Table 2: ROUGE F-score and sentiment correlation for optimal values of α with baseline 6

proposed functions not only outperform the baselines in terms of ROUGE scores for optimal parameters but also, give better correlation with the document sentiment. This can be quantitively verified by test of significance, unpaired one-tailed t-test without assuming equal variance between the baselines and the systems. The p-values are 0.0203 and 0.0066 respectively for ROUGE-1 F Score and Sentiment Correlation, justifying that the performance improvement by our system over the baselines is statistically significant at $p < 0.05$. The main reason being that the functions with optimal values of trade-off parameter α strike out a balance between relevance and subjectivity. Clearly, the facility location based monotone submodular functions are the best choice as objective for opinion summarization task as they select sentences with maximum subjectivity (facilities giving best service).

Our system is able to access the information of aspect and polarity of each sentence, while some baselines do not. So, the improvement over the baselines may be attributed to those additional information rather than the optimality of the partial enumeration greedy algorithm over submodular functions. So, we therefore, introduced the following baseline to question this misdoubt on the experiment:

6. **Baseline-6/LIN** :

In this baseline, the greedy algorithm (Lin and Bilmes, 2010) is used for summary generation, using the same functions and information in the formulation. This algorithm fills the empty summary set greedily by adding a single sentence in each

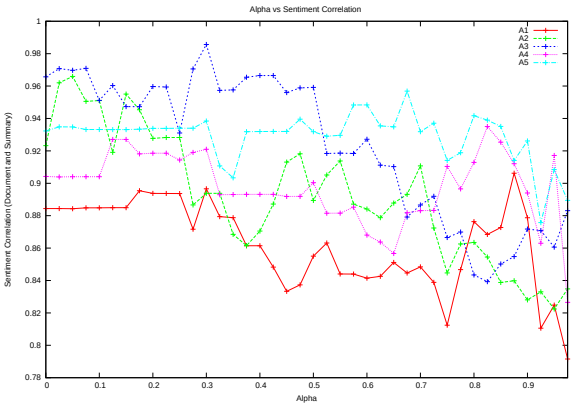

Figure 2: Sentiment Correlation *vs* α

iteration, which gives maximum return over cost ratio ($\frac{F(S^*)-F(S)}{len(s)}$), ensuring that current solution is feasible with respect to the knapsack constraint (cost($S \cup \{newSentence\}$) $\leq$ B). This algorithm has a complexity of O(n^2) but gives only performance guarantee of $\frac{1}{2}(1 - e^1) \approx 0.316$ (Khuller et al., 1999).

Table 2 compares the same five functions in our system with (Lin and Bilmes, 2010) system based on optimal values of tradeoff α. From the table, it can be inferred that our system also outperforms this baseline both in terms of ROUGE scores and sentiment correlation, which can be quantitively verified by test of significance, unpaired one-tailed t-test without assuming equal variance between the baselines and the systems. The p-values are 0.02517 and 0.003965 respectively for ROUGE-1 F Score and Sentiment Correlation, justifying that the performance improvement by our system over LIN system is statistically significant at $p < 0.05$.

The figures 2 and 3 plot the value of senti-

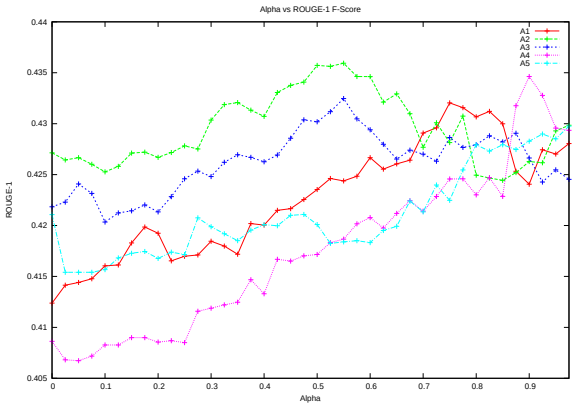

Figure 3: ROUGE-1 F-score *vs* α

Sys	ROUGE1	ROUGE2	Senti. Corr.
A_1	0.43223	0.15702	0.84827
A_2	0.43594	**0.15977**	0.88601
A_3	0.43247	0.15436	0.87038
A_4	**0.43602**	0.15760	0.87818
A_5	0.42976	0.15551	**0.90147**

Table 3: Maximum ROUGE F-score and their corresponding sentiment correlation

ment correlation and ROUGE-1 F score for the formulated submodular functions with respect to the trade-off parameter, α respectively. Looking at the graph 2, we can observe that more weightage to relevance over subjective coverage of aspects decreases the sentiment correlation, which was expected because the summary generated misses out on subjective sentiment due to trade-off. Similarly, by looking at the graph 3, we also observe that more weightage to relevance over subjective coverage of aspects increases the ROUGE score as expected. The erratic behaviour in figure 3 can be explained by arguing that subjective words are also important for summary and thus, giving less weightage to them over relevance, ROUGE score will increase but not properly.

The table 3 presents the value of sentiment correlation corresponding to maximum ROUGE score (for $\alpha \approx 1$). Clearly, A_4 and A_2 have maximum ROUGE scores as they neglect polarities and instead, reward on aspect based partitions, thus increasing coverage. The table 4 presents the value of ROUGE score corresponding to maximum sentiment correlation (for $\alpha \approx 0$). Clearly, A_4 also has maximum sentiment correlation as it rewards maximum subjectivity, irrespective of polarities and

Sys	Senti. Corr.	ROUGE1	ROUGE2
A_1	0.95306	**0.42572**	0.14939
A_2	0.97538	0.41764	0.14836
A_3	0.93155	0.42415	0.14782
A_4	**0.98566**	0.42492	**0.14942**
A_5	0.95415	**0.42572**	0.14266

Table 4: Maximum sentiment correlation and corresponding ROUGE F-Score

the corresponding ROUGE-2 F-score is also highest among all functions. Tables 1 and 2 contain the ROUGE F-score and sentiment correlation for optimal values of α, found after grid search while tables 3 and 4 contain the peak values in the figures 2 and 3. For example, table 3 contains the peak value of ROUGE-1 F score from figure 3 and the corresponding value of Sentiment Correlation from figure 2, at the same α.

7 Conclusion

In this paper, we show that conflict between subjectivity and relevance naturally arises in opinion summarization. To address this problem, we introduce new monotone submodular functions that are well suited to document summarization (Lin and Bilmes, 2010; Lin and Bilmes, 2011; Morita et al., 2013) by modeling two important properties of opinion summary - relevance and subjective coverage of aspects. We then, design different possible combinations of objective functions to model the task. To solve the algorithm effectively, we use the partial enumeration based algorithm, which is though computationally expensive (O(n^5) function calls), gives a performance guarantee of 63% for an NP-hard problem like summarization (McDonald, 2007). We have justified the submodular property of opinion summary through examples and significant performance of the system over the baselines. Further, this optimal trade-off between relevance and subjectivity can be used to design an evaluation framework for opinion summarization task as both part of the objective functions are proportional to the ROUGE and Sentiment Correlation respectively, which are widely used evaluation measures (Kim et al., 2011). As opinion summarization task lies in the intersection of opinion mining and summarization problems, both IR and NLP communities will benefit from our work.

References

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, pages 335–336.

Andrea Esuli and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422.

Elena Filatova. 2004. Event-based extractive summarization. In *Proceedings of ACL Workshop on Summarization*, pages 104–111.

Samir Khuller, Anna Moss, and Joseph Seffi Naor. 1999. The budgeted maximum coverage problem. *Information Processing Letters*, 70(1):39–45.

Hyun Duk Kim, Kavita Ganesan, Parikshit Sondhi, and ChengXiang Zhai. 2011. Comprehensive review of opinion summarization.

Andreas Krause and Daniel Golovin. 2014. Submodular function maximization. In *Tractability: Practical Approaches to Hard Problems (to appear)*. Cambridge University Press, February.

Kevin Lerman, Sasha Blair-Goldensohn, and Ryan McDonald. 2009. Sentiment summarization: evaluating and learning user preferences. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 514–522.

Hui Lin and Jeff Bilmes. 2010. Multi-document summarization via budgeted maximization of submodular functions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 912–920.

Hui Lin and Jeff Bilmes. 2011. A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 510–520.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Ryan McDonald. 2007. A study of global inference algorithms in multi-document summarization. In *Advances in Information Retrieval*, pages 557–564. Springer.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into texts. In *Proceedings of EMNLP*, volume 4, page 275.

Hajime Morita, Hiroya Takamura, Ryohei Sasano, and Manabu Okumura. 2013. Subtree extractive summarization via submodular maximization. In *Proceedings of 51st Annual Meeting of the Association for Computational Linguistics (to appear)*.

Hitoshi Nishikawa, Takaaki Hasegawa, Yoshihiro Matsuo, and Genichiro Kikui. 2010a. Opinion summarization with integer linear programming formulation for sentence extraction and ordering. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 910–918.

Hitoshi Nishikawa, Takaaki Hasegawa, Yoshihiro Matsuo, and Genichiro Kikui. 2010b. Optimizing informativeness and readability for sentiment summarization. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 325–330.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Pavel Pecina and Pavel Schlesinger. 2006. Combining association measures for collocation extraction. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 651–658. Association for Computational Linguistics.

Korbinian Riedhammer, Benoit Favre, and Dilek Hakkani-Tür. 2010. Long story short–global unsupervised models for keyphrase based meeting summarization. *Speech Communication*, 52(10):801–815.

Maxim Sviridenko. 2004. A note on maximizing a submodular set function subject to a knapsack constraint. *Operations Research Letters*, 32(1):41–43.

Laurence A Wolsey. 1982. An analysis of the greedy algorithm for the submodular set covering problem. *Combinatorica*, 2(4):385–393.