Summarizing Contrastive Viewpoints in Opinionated Text

Michael J. Paul[*]
University of Illinois
Urbana, IL 61801, USA
mjpaul2@illinois.edu

ChengXiang Zhai
University of Illinois
Urbana, IL 61801, USA
czhai@cs.uiuc.edu

Roxana Girju
University of Illinois
Urbana, IL 61801, USA
girju@illinois.edu

Abstract

This paper presents a two-stage approach to summarizing multiple contrastive viewpoints in opinionated text. In the first stage, we use an unsupervised probabilistic approach to model and extract multiple viewpoints in text. We experiment with a variety of lexical and syntactic features, yielding significant performance gains over bag-of-words feature sets. In the second stage, we introduce *Comparative LexRank*, a novel random walk formulation to score sentences and pairs of sentences from opposite viewpoints based on both their representativeness of the collection as well as their contrastiveness with each other. Experimental results show that the proposed approach can generate informative summaries of viewpoints in opinionated text.

1 Introduction

The amount of opinionated text available online has been growing rapidly, increasing the need for systems that can summarize opinions expressed in such text so that a user can easily digest them. In this paper, we study how to summarize opinionated text in a such a way that highlights contrast between multiple viewpoints, which is a little-studied task.

Usually, online opinionated text is generated by multiple people, and thus often contains multiple viewpoints about an issue or topic. A viewpoint/perspective refers to "a mental position from which things are viewed" (cf. WordNet). An opinion is usually expressed in association with a particular viewpoint, even though the viewpoint is usually not explicitly given; for example, a blogger that is in favor of a policy would likely look at the positive aspects of the policy (i.e., positive viewpoint), while someone against the policy would likely emphasize the negative aspects (i.e., negative viewpoint). Moreover, in an opinionated text with diverse opinions, the multiple viewpoints taken by opinion holders are often "contrastive", leading to opposite polarities. Indeed, such contrast in opinions may be a main driving force behind many online discussions.

Futhermore, opinions regarding news events and other short-term issues may quickly emerge and disappear. Such opinions may reflect many different types of viewpoints which cannot be modeled by current systems. For this reason, we believe that a viewpoint summarization system would benefit from the ability to extract unlabeled viewpoints without supervision. Even if such clustering has inaccuracies, it could still be a useful starting point for human editors to select representative excerpts.

Thus, given a set of opinionated documents about a topic, we aim at automatically extracting and summarizing the multiple contrastive viewpoints implicitly expressed in the opinionated text to facilitate digestion and comparison of different viewpoints. Specifically, we will generate two types of multi-view summaries: macro multi-view summary and micro multi-view summary. A macro multi-view summary would contain multiple sets of sentences, each representing a different viewpoint; these different sets of sentences can be compared to understand the difference of multiple viewpoints at the "macro level." A micro multi-view summary would contain a set of pairs of contrastive sentences (each pair

[*] Now at Johns Hopkins University (mpaul@cs.jhu.edu).

Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 66–76,
MIT, Massachusetts, USA, 9-11 October 2010. ©2010 Association for Computational Linguistics

consists of two sentences representing two different viewpoints), making it easy to understand the difference between two viewpoints at the "micro level."

Although opinion summarization has been extensively studied (e.g., (Liu et al., 2005; Hu and Liu, 2004; Hu and Liu, 2006; Zhuang et al., 2006)), existing work has not attempted to generate our envisioned contrastive macro and micro multi-view summaries in an unsupervised way, which is the goal of our work. For example, Hu and Liu (2006) rank sentences based on their dominant sentiment according to the polarity of adjectives occuring near a product feature in a sentence. A contradiction occurs when two sentences are highly unlikely to be simultaneously true (cf. (Marneffe et al., 2008)). Although little work has been done on contradiction detection, there are a few notable approaches (Harabagiu et al., 2006; Marneffe et al., 2008; Kim and Zhai, 2009).

The closest work to ours is perhaps that of Lerman and McDonald (2009) who present an approach to contrastive summarization. They add an objective to their summarization model such that the summary model for one set of text is different from the model for the other set. The idea is to highlight the key differences between the sets, however this is a different type of contrast than the one we study here – our goal is instead to make the summaries *similar* to each other, to contrast how the same information is conveyed through different viewpoints.

In this paper, we propose a two-stage approach to solving this novel summarization problem, which will be explained in the following two sections.

2 Modeling Viewpoints

The first challenge to be solved in order to generate a contrastive summary of multiple viewpoints is to model and extract these viewpoints which are hidden in text. In this paper we propose to solve this challenge by employing the Topic-Aspect Model (TAM) (Paul and Girju, 2010), which is an extension of the Latent Dirichlet Allocation (LDA) model (Blei et al., 2003) for jointly modeling topics and viewpoints in text. While most existing work on such topic models (including TAM) has taken a topic model as a generative model for word tokens in text, we propose to take TAM as a generative model for more complex linguistic features extracted from text. These are more discriminative than single word tokens and can improve the accuracy of extracting multiple viewpoints as we will show in the experimental results' section. Below we first give a brief introduction to TAM and then present the proposed set of features.

2.1 Topic-Aspect Model (TAM)

LDA-style probabilistic topic models of document content (Blei et al., 2003) have been shown to offer state-of-the-art summarization quality. Such models also provide a framework for adding additional structure to a summarization model (Haghighi and Vanderwende, 2009). In our case, we want to add more structure to a model to incorporate the notion of viewpoint/perspective into our summaries.

When it comes to extracting viewpoints, recent research suggests that it may be beneficial to model both topics and perspectives, as sentiment may be expressed differently depending on the issue involved (Brody and Elhadad, 2010; Paul and Girju, 2010). For example, let's consider a set of product reviews for a home theater system. Content topics in this data might include things like sound quality, usability, etc., while the viewpoints might be the positive and negative sentiments. A word like *speakers*, for instance depends on the sound topic but not a viewpoint, while *good* would be an example of a word that depends on a viewpoint but not any particular topic. A word like *loud* would depend on both (since it would be considered positive sentiment only in the context of the sound quality topic), while a word like *think* depends on neither.

We make use of a recent model, the Topic-Aspect Model (Paul and Girju, 2010), which can model such behavior with or without supervision. Under this model, a document has a mixture over topics as well as a mixture over viewpoints. The two mixtures are drawn independently of each other, and thus can be thought of as two separate clustering dimensions. A word is associated with variables denoting its topic and viewpoint assignments, as well as two binary variables to denote if the word depends on the topic and if the word depends on the viewpoint. A word may depend on the topic, the viewpoint, both, or neither, as in the above example.

The generative process for a document d under this model can be briefly described as follows. For each word in a document:

1. Sample a topic z from $P(z|d)$ and a viewpoint v from $P(v|d)$.
2. Sample a "level" $\ell \in \{0, 1\}$ from $P(\ell|d)$. This determines if the word will depend on the topic (topical level) or not (background level).
3. Sample a "route" $r \in \{0, 1\}$ from $P(r|\ell, z)$. This determines if the word will depend on the viewpoint.
4. Sample a word w from $P(w|z, v, r, \ell)$.

The probabilities are multinomial/binomial distributions with Dirichlet/Beta priors, and thus this model falls under the standard LDA framework. The number of topics and number of viewpoints are parameters that must be specified. Inference can be done with Gibbs sampling (Paul and Girju, 2010).

TAM naturally gives us a very rich output to use in a viewpoint summarization application. If we are doing unsupervised viewpoint extraction, we can use the output of the model to compute $P(v|sentence)$ which could be used to generate summaries that contain only excerpts that strongly highlight one viewpoint over another. Similarly, we could use the learned topic mixtures to generate topic-specific summaries. Futhermore, the variables r and ℓ tell us if a word is dependent on the viewpoint and topic, and we could use this information to focus on sentences that contain informative content words. Note that without supervision, TAM's clustering is based only on co-occurrences and the patterns it captures may or may not correspond with the viewpoints we wish to extract. Nonetheless, we show in this research that it can indeed find meaningful viewpoints with reasonable accuracy on certain data sets. Although we do not explore this in this paper, additional information about the viewpoints could be added to TAM by defining priors on the distributions to further improve the accuracy of viewpoint discovery.

2.2 Features

Previous work with TAM used only bag of words features, which may not be the best features for capturing viewpoints. For example, "Israel attacked Palestine" and "Palestine attacked Israel" are identical excerpts in an exchangable bag of words representation, yet one is more likely to come from the perspective of a Palestinian and the other from an Israeli. In this subsection, we will propose a variety of feature sets. We evaluate the utility of these features to the task of modeling viewpoints by measuring the accuracy of unsupervised clustering.

2.2.1 Words

We have experimented with simple bag of words features as baseline approaches, both with and without removing stop words, and found that the accuracy of clustering by viewpoint is better when retaining all words. This supports the observation that common function words may have important psychological properties (Chung and Pennebaker, 2007). Thus, we do not do any stop word removal for any of our other feature sets. We find that we get better results by stemming the words, so we apply Porter's stemmer to all of our features described.

2.2.2 Dependency Relations

It has been shown that using syntactic information can improve the accuracy of sentiment models (Joshi and Rosé, 2009). Thus, instead of representing documents as a bag of words, we will experiment with using features returned by a dependency parser. For this, we used the Stanford parser[1], which returns dependency tuples of the form $\texttt{rel}(a, b)$ where $\texttt{rel}$ is some dependency relation and a and b are tokens of a sentence. We can use these specific tuples as features, referred here as the *full-tuple* representation.

One problem with this representation is that we are using very specific information and it is harder for learning algorithms to find patterns due to the lack of redundancy. One solution is to generalize these features and rewrite a tuple $\texttt{rel}(a, b)$ as two tuples: $\texttt{rel}(a, *)$ and $\texttt{rel}(*, b)$ (Greene and Resnik, 2009; Joshi and Rosé, 2009). We will refer to this as the *split-tuple* representation.

2.2.3 Negation

If a word w_i appears in the head of a $\texttt{neg}$ relation, then we would like this to be reflected in other dependency tuples in which w_i occurs. For a tuple $\texttt{rel}(w_i, w_j)$, if either w_i or w_j is negated, then we simply rewrite it as $\neg\texttt{rel}(w_i, w_j)$.

An alternative would be to rewrite the individual word w_i as $\neg w_i$. However in our experiments this representation produced worse accuracies, perhaps because this produces less redundancy.

[1] http://nlp.stanford.edu/software/

2.2.4 Polarity

We also hypothesize that lexical polarity information may improve our model. If we are using the *full-tuple* representation, then a tuple becomes more general by replacing the specific word with a $+$ or $-$. In the case that both words are polarity words, we use two tuples, replacing only one word at a time rather than replacing both words with their polarity signs. To determine the polarity of a word, we simply use the Subjectivity Clues lexicon (Wilson et al., 2005) and as polarity values, *positive* (+), *negative* (-), and *neutral* (*). Under our *split-tuple* representation, this becomes more specific by replacing the $*$ with the polarity sign. For example, the tuple $\texttt{amod}(idea, good)$ would be represented as $\texttt{amod}(idea, +)$ and $\texttt{amod}(*, good)$. We collapse negated features to flip the polarity sign such that $\neg\texttt{rel}(a, +)$ becomes $\texttt{rel}(a, -)$.

2.2.5 Generalized Relations

We also experimented with backing off the relations themselves. Since the Stanford dependencies can be organized in a hierarchy[2], we will represent the relations at more generalized levels in the hierarchy. For example, both a direct object and an indirect object are a type of object. For a relation rel, we define $\texttt{R}_{\texttt{rel}}$ as the relation above rel in the hierarchy – for example, $\texttt{R}_{\texttt{dobj}} = \texttt{obj}$. We make an exception for neg which has its own important properties that we wish to retain, so we let $\texttt{R}_{\texttt{neg}} = \texttt{neg}$. Thus, when using these features, we rewrite $\texttt{rel}(a, b)$ as $\texttt{R}_{\texttt{rel}}(a, b)$.

3 Multi-Viewpoint Summarization

As a computation problem, extractive multi-viewpoint summarization would take as input a set of candidate excerpts[3] $X = \{x_1, x_2, ..., x_{|X|}\}$ with k viewpoints and generate two types of multi-view contrastive summaries: 1) A macro contrastive summary S_{macro} consists of k disjoint sets of excerpts, $X_1, X_2, ..., X_k \subset X$ with each X_i containing representative sentences of the i-th view (i.e., $S_{macro} = (X_1, ..., X_k)$). The number of excerpts in each X_i can be empirically set based on application needs. 2) A micro contrastive summary S_{micro} consists of a set of excerpt pairs, each containing two excerpts from two different viewpoints, i.e., $S_{micro} = \{(s_1, t_1), ..., (s_n, t_n)\}$ where $s_i \in X$ and $t_i \in X$ are two comparable excerpts representing two different viewpoints. n is the length of the summary, which can be set empirically based on application needs. Note that both macro and micro summaries can reveal contrast between different viewpoints, though at different granularity levels.

To generate macro and micro summaries based on the probabilistic assignment of excerpts to viewpoints given by TAM, we propose a novel extension to the LexRank algorithm (Erkan and Radev, 2004), a graph-based method for scoring representative excerpts to be used in a summary. Our key idea is to modify the definition of the jumping probability in the random walk model so that it would favor excerpts that represent a viewpoint well and encourage jumping to an excerpt comparable with the current one but from a different viewpoint. As a result, the stationary distribution of the random walk model would capture representative contrastive excerpts and allow us to generate both macro and micro contrastive summaries within a unified framework. We now describe this novel summarization algorithm (called Comparative LexRank) in detail.

3.1 Comparative LexRank

LexRank is a PageRank-like algorithm (Page et al., 1998), where we define a random walk model on top of a graph that has sentences to be summarized as nodes and edges placed between two sentences that are similar to each other. We can then score all the sentences based on the expected probability of a random walker visiting each sentence. We use the short-hand $P(x_j|x_i)$ to denote the probability of being at node x_j at a time t given that the walker was at x_i at time $t - 1$. The jumping probability from node x_i to node x_j is given by:

$$P(x_j|x_i) = \frac{sim(x_i, x_j)}{\sum_{j' \in X} sim(x_i, x_{j'})} \tag{1}$$

where sim is a content similarity function defined on two sentence/excerpt nodes.

Our extension is mainly to modify this jumping probability in two ways so as to favor visiting contrastive representative opinions from multiple view-

[2] The complete hierarchy can be found in the Stanford dependencies manual (Marneffe and Manning, 2008).

[3] An "excerpt" refers to the smallest unit of text that will make up our summary such as a sentence.

points. The first modification is to make it favor jumping to a good representative excerpt x of any viewpoint v (i.e., with high probability $p(v|x)$ according to the TAM model). The second modification is to further favor jumping between two excerpts that can potentially form a good contrastive pair for use in generating a micro contrastive summary.

Specifically, under our model, the random walker first decides whether to jump to a sentence of the same viewpoint or to a sentence of a different viewpoint. We define this decision as a binary variable $z \in \{0, 1\}$. Intuitively, if we can force the random walker to move back and forth between viewpoints, then the final scores will favor sentences that are similar across both viewpoints.

We define two different modified similarity functions for the two possible values of z. The first one, sim_0 (corresponding to $z = 0$) scales the similarity by the likelihood that the two x's represent the same viewpoint, and the second one, sim_1 (for $z = 1$) scales the similarity by the likelihood that the x's come from different viewpoints.

$$\begin{aligned} sim_0(x_i, x_j) &= sim(x_i, x_j) \sum_{m=1}^{k} P(v = m|x_i)P(v = m|x_j) \\ sim_1(x_i, x_j) &= sim(x_i, x_j) \times \\ &\sum_{m_1, m_2 \in [1,k], m_1 \neq m_2} P(v = m_1|x_i)P(v = m_2|x_j) \end{aligned}$$

where $P(v|x)$ denotes the probability that the excerpt x belongs to the viewpoint v, and in general, can be obtained through any multi-viewpoint model. A special case of this is when the labels for viewpoints are known, in which case $P(v|x) = 1$ for the correct label and 0 for the others.

In our experiments, $P(v|x)$ comes from the output of TAM, and we define $sim(x_i, x_j)$ as the cosine between the vectors x_i and x_j, although again any similarity function could be used. The conditional transition probability from x_i to x_j given z is then:

$$P(x_j|x_i, z) = \frac{sim_z(x_i, x_j)}{\sum_{j' \in X} sim_z(x_i, x_{j'})} \quad (2)$$

Using λ to denote $P(z = 0)$ and marginalizing across z, we have the transition probability:

$$P(x_j|x_i) = \lambda P(x_j|x_i, z = 0) + (1 - \lambda)P(x_j|x_i, z = 1)$$

The stationary distribution of the random walk gives us a scoring of the excerpts to be used in our summary. It is also possible to score *pairs* of excerpts that contrast each other. We define the score for a pair (x_i, x_j) as the probability of being at x_i and transitioning to x_j or vice versa, where x_i and x_j are of opposite viewpoints. Specifically:

$$P(x_i)P(x_j|x_i, z = 1) + P(x_j)P(x_i|x_j, z = 1) \quad (3)$$

3.2 Summary Generation

The final summary should be a set of excerpts that have a high relevance score according to our scoring algorithm, but are not redundant among each other. Many techniques could be used to accomplish this (Carbonell and Goldstein, 1998; McDonald, 2007), but we use a simple greedy approach: at each step of the summary generation algorithm, we add the excerpt with the highest relevance score as long as the excerpt's redundancy score – the cosine similarity between the candidate and the current summary – is under some threshold δ. This is repeated until the summary reaches a user-supplied length limit.

Macro contrastive summarization: A macro-level summary consists of independent summaries for each viewpoint, which we generate by first using the random walk stationary distribution across all of the data to rank the excerpts. We then separate the top-ranked excerpts into two disjoint sets according to their viewpoint based on whichever gives a greater value of $P(v|x)$, and finally remove redundancy and produce the summary according to our method described above. We refer to this as *macro contrastive summarization*, because the summaries will contrast each other in that they have related content, but the excerpts in the summaries are not explicitly aligned with each other.

Micro contrastive summarization: A candidate excerpt for a micro-level summary will consist of a pair (x_i, x_j) with the pairwise relevance score defined in Equation 3. We can then rank these pairs and remove redundancy. It is possible that both x_i and x_j in a high-scoring pair may belong to the same viewpoint; such a case would be filtered out since we are mainly interested in including contrastive pairs in our summary. We refer to this as *micro contrastive summarization*, because the summaries will allow us to see contrast at the level of individual excerpts from different viewpoints.

4 Experiments and Evaluation

4.1 Experimental Setup

Evaluation of multi-view summarization is challenging as there is no existing data set we can use. We leverage the resources on the Web and created two data sets in the domain of political opinion.

Our first dataset is a set of 948 verbatim responses to a Gallup® phone survey about the 2010 U.S. healthcare bill (Jones, 2010), conducted March 4-7, 2010. Responses in this set tend to be short and often incomplete or otherwise ill-formed and informal sentences. Respondants indicate if they are 'for' or 'against' the bill, and there is a roughly even mix of the two viewpoints (45% for and 48% against).

We also use the Bitterlemons corpus, a collection of 594 editorials about the Israel-Palestine conflict. This dataset is fully described in (Lin et al., 2006) and has been used in other perspective modeling literature (Lin et al., 2008; Greene and Resnik, 2009). The style of this data differs substantially from the healthcare data in that documents in this set tend to be long and verbose articles with well-formed sentences. It again contains a fairly even mixture of two different perspectives: 312 articles from Israeli authors and 282 articles from Palestinian authors.

Moreover, for the healthcare data set, manually extracted opinion polls are available on the Web, which we further leverage to construct gold standard summaries to evaluate our method quantitatively. The data and test sets are available at `http://apfel.ai.uiuc.edu/resources.html`.

4.2 Stage One: Modeling Viewpoints

The main research question we want to answer in modeling viewpoints is whether richer feature sets would lead to better accuracy than word features. We used our various feature sets as input to TAM and measured the accuracy of clustering documents by viewpoint. This evaluation serves both to measure how accurately this type of clustering can be done, as well as to measure which types of features are important for modeling viewpoints.

We found that the clustering accuracy is improved if we measure the accuracy of only the subset of documents such that $P(v|doc)$ is greater than some threshold (we used 0.8). Thus, the accuraries presented in this section are measured using this confidence threshold. We will use this approach for the summarization task as well, as it ensures we are only summarizing documents where we have high confidence about their viewpoint membership.

There are several parameters to set for TAM. Since our focus is on comparing linguistic features with word features, we simply set these parameters to some reasonable values: We used Dirichlet pseudo-counts of 80.0 for $P(\ell = 0)$, 20.0 for $P(\ell = 1)$, uniform pseudo-counts of 5.0 for $P(x)$, 0.1 for the topic and aspect mixtures, and 0.01 for the word distributions. We tell the model to use 2 viewpoints as well as 5 topics for the healthcare corpus and 8 topics for the Bitterlemons corpus.

There is high variance in the accuracies depending on how the Gibbs samplers were initialized. We thus repeated the experiments many times to obtain relatively confident measures – 200 times for the healthcare set and 50 times for the Bitterlemons set, with 2000 iterations each time. A natural way to select a model is to choose the model that gives the highest likelihood to its input. To evaluate how well this selection strategy would work, we measured the correlation between accuracy and likelihood.

The results are shown in Table 1. We can make several observations. (1) In all cases, the proposed linguistic features yield higher accuracy than the word features, supporting our hypothesis that for viewpoint modeling, applying TAM to these features improves performance over using simple word features. Since virtually all existing work on topic models assumes word tokens as data to be modeled, our results suggest that it would be interesting to explore applying generative topic models to complex features for other tasks as well. This may be because by adding additional complex features to the observed data, we artificially inflate the data likelihood to emphasize modeling co-occurrences of such features, which effectively biases the model to capture a certain perspective of co-occurrences.

(2) The increase is substantially greater for the Bitterlemons corpus, which may be due to the fact that the parsing accuracy is likely better because the language is formal. The *split-tuple* representation is very significantly better for the healthcare corpus, but it is not clear which is better for the Bitterlemons corpus. It is also not clear how the generalized relations affect the performance.

	Healthcare Corpus					*Bitterlemons Corpus*				
Feature Set	**Mean**	**Med**	**Max**	**MaxLL**	**Corr**	**Mean**	**Med**	**Max**	**MaxLL**	**Corr**
bag of words	61.12 +/- 0.76%	61.01	72.17	52.92	0.187	68.22 +/- 3.31%	69.26	88.27	84.94	0.39
- no stopwords	60.58 +/- 0.79%	60.50	72.18	62.58	0.154	61.29 +/- 3.05%	57.69	91.34	82.91	0.33
full-tuples	62.42 +/- 0.88%	62.47	74.04	63.37	0.201	80.89 +/- 3.45%	85.40	94.07	92.10	0.34
+ negation	63.67 +/- 0.81%	64.54	74.07	69.25	0.338	80.60 +/- 3.88%	88.07	95.61	91.32	0.66
+ neg. + polarity	63.16 +/- 0.94%	64.46	74.05	67.8	0.455	82.53 +/- 3.55%	86.64	94.44	91.16	0.31
gen. full-tuples	63.80 +/- 0.73%	64.35	73.29	71.70	0.254	76.62 +/- 4.09%	84.56	94.53	84.56	0.25
split-tuples	68.32 +/- 0.90%	70.74	77.80	76.57	0.646	77.14 +/- 3.64%	81.29	92.99	88.13	0.30
+ negation	68.00 +/- 0.91%	69.11	79.73	76.14	0.187	83.53 +/- 3.05%	87.71	95.00	95.00	0.12
+ neg. + polarity	65.11 +/- 1.05%	65.35	78.59	67.22	0.159	81.24 +/- 3.37%	83.44	95.03	88.55	0.08
gen. split-tuples	69.31 +/- 0.83%	70.69	77.90	73.90	0.653	76.69 +/- 4.36%	83.78	93.60	91.67	0.09

Table 1: The clustering accuracy with TAM using a variety of feature sets. These results were averaged over 200 randomly-initialized Gibbs sampling procedures for the healthcare set, and 50 procedures for the Bitterlemons set. The 95% confidence interval using a standard t-test is also given. *Max* refers to the maximum accuracy obtained over the 200 or 50 instances. *MaxLL* refers to the clustering accuracy using the model that yielded the highest corpus log-likelihood as defined by TAM. *Corr* refers to the Pearson correlation coefficient between accuracy and log-likelihood.

(3) It appears that adding polarity helps the *full-tuple* features (by making them more general) but hurts the *split-tuple* features (by making them more specific). Negation significantly improves the *full-tuple* features in the Bitterlemons corpus, but it is not clear if it helps in the other cases. It should be noted that capturing negation and polarity is a very complex and difficult task, and it is not expected that our simple approaches will accurately capture these properties. Nonetheless, it seems that these simple features may help in certain cases.

4.3 Stage Two: Summarizing Viewpoints

For the second stage (i.e., the Comparative LexRank algorithm), we mainly want to evaluate the quality of the generated contrastive multi-viewpoint summary and study the effectiveness of our extension to the standard LexRank. Below we present extensive evaluation of our summarization method on the healthcare data. We do not have an evaluation set with which to compute quantitative metrics on the Bitterlemons corpus, so we will instead perform a simple qualitative evaluation in the last subsection.

4.3.1 Gold Standard Summaries

The responses to the Gallup healthcare poll are described in an article[4] which gives a table of the main responses found in the data along with their prominence in the data. In a way, this represents an expert human-generated summary of our database, and we will use this as a gold standard macro contrastive summary against which the representativeness of a multi-viewpoint contrastive summary can be evaluated. The reasons given in this table will be used verbatim as our reference set, excluding the other/no-reason/no-opinion reasons. A sample of this table is shown in Table 2.

We also want to develop a reference set for micro contrastive summaries, where we are mainly interested in evaluating contrastiveness. To do this, we asked 3 annotators to identify contrastive pairs in the "main reasons" table described above. Each pair must contain one reason from the 'for' side and one reason from the 'against' side, though we do not require a one-to-one alignment; that is, multiple pairs may contain the same reason. We take the set of pairs that were identified as being contrastive by at least 2 annotators to be our gold set of contrastive pairs. Because these pairs come from the gold summary, they are still representative of the collection as a whole, rather than fine-grained contrasts.

The macro reference set contains 9 'for' reasons and 15 'against' reasons. The micro reference set contains 13 annotator-identified pairs composed of 9 unique 'for' reasons and 8 unique 'against' reasons.

4.3.2 Baseline Approaches

Graph-based algorithms: The standard LexRank algorithm can also be used to score pairs of sentences according to Equation 3. We will thus compare our new LexRank extension to the unmodified form of this algorithm. When $\lambda = 1$, the random walk model only transitions to sentences within the same viewpoint, and thus in this case our modified algorithm produces the same ranking as the unmodified LexRank. This will be our first baseline.

[4] http://www.gallup.com/poll/126521/Favor-Oppose-Obama-Healthcare-Plan.aspx

For		*Against*	
People need health insurance/Too many uninsured	29%	Will raise costs of insurance/Make it less affordable	20%
System is broken/Needs to be fixed	18%	Does not address real problems	19%
Costs are out of control/Would help control costs	12%	Need more information/clarity on how system would work	8%
Moral responsibility to provide/Obligation/Fair	12%	Against big government/Too much government involvement	8%

Table 2: Some of the top reasons given along with their prominence in the healthcare data, as analyzed by Gallup. This is a sample of what will serve as our gold set. The highlighted cells show an example of a contrastive pair identified by our annotators.

Model-based algorithms: We will also compare against the approach of Lerman and McDonald (2009) who introduce their contrastiveness objective into a model-based summarization algorithm. The basic form of this algorithm is to select a set of sentences S_m to minimize the KL-divergence between the models of the summary S_m and the entire collection X_m for a viewpoint m. The objective function is: $-\sum_{m=1}^{k} KL(L(S_m)||L(X_m))$ where L is an arbitrary language model. We define $L(A)$ simply as the unigram distribution over words in the collection A, a method also evaluated by Haghighi and Vanderwende (2009). This is the fairest comparison to our LexRank experiments, where sentences are also represented as unigrams. (We do not do any modeling with TAM in our quantitative evaluation.)

Lerman and McDonald introduce an additional term to maximize the KL-divergence between the summary of one viewpoint and the collection of the opposite viewpoint, so that each viewpoint's summary is dissimilar to the other viewpoints. We borrow this idea but instead do the opposite so that the viewpoints' summaries are more (rather than less) similar to each other. This contrastive version of our model-based baseline is formulated as:

$$-\sum_{m_1=1}^{k} KL(L(S_{m_1})||L(X_{m_1})) + \Big(\tfrac{1}{k-1}\textstyle\sum_{m_2\in[1,k],m_1\neq m_2} KL(L(S_{m_1})||L(X_{m_2}))\Big)$$

Our summary generation algorithm is to iteratively add excerpts to the summary in a greedy fashion, selecting the excerpt with the highest score in each iteration. Note that this approach only generates macro-level summaries, leaving us with the LexRank baseline for micro-level summaries.

4.3.3 Metrics

We will evaluate our summaries using a variant of the standard ROUGE evaluation metric (Lin, 2004).

Recall that we have two different evaluation sets – one that contains all of the reasons for each viewpoint, and one that consists only of aligned pairs of excerpts. Since the same excerpt may appear in multiple pairs, there would be significant redundancy in our reference summary if we were to include every pair. Thus, we will restrict a contrastive reference summary to exclude overlapping pairs, and we will have many reference sets for all possible combinations of pairs. There is only one reference set for the representativeness criterion.

Our reference summaries have a unique property in that the summaries have already been annotated with the prominence of the different reasons in the data. A good summary should capture the more prominent statements, so we will include this in our scoring function. We thus augment the basic ROUGE n-gram recall score by weighting the n-gram counts in the reference summary according to this percentage. This is a generalization of the standard ROUGE formula where this percentage would be uniform.

For evaluating the macro-level summaries, we will score the summaries for the two viewpoints separately, given a reference set Ref_i and a candidate summary C_i for a viewpoint $v = i$. The final score is a combination of the scores for both viewpoints, i.e. $S_{rep} = 0.5S(Ref_i, C_i) + 0.5S(Ref_j, C_j)$ where $S(Ref, C)$ is our ROUGE-based scoring metric. It would also be interesting to measure how well a viewpoint's summary matches the gold summary of the *opposite* viewpoint, which will give insights into how well the Comparative LexRank algorithm makes the two summaries similar to each other. We will measure this as the inverse of the above metric, i.e. $S_{opp} = 0.5S(Ref_i, C_j) + 0.5S(Ref_j, C_i)$.

Finally, to score the micro-level comparative summaries (recall that this gives explicitly-aligned pairs of excerpts), we will concatenate each pair (x_i, x_j) as a single excerpt, and use these as the excerpts in our reference and candidate summaries. The scoring function is then $S_p = S(Ref_{pairs}, C_{pairs})$. Note that we have multiple reference summaries for the

λ	S_{rep}**-1**	S_{opp}**-1**	S_{rep}**-2**	S_{opp}**-2**	S_p**-1**	S_p**-2**
0.0	**.425**	**.416**	.083	.060	.309	.036
0.2	**.410**	**.423**	.082	**.065**	.285	.044
0.5	**.419**	**.434**	.085	**.072**	**.386**	.044
0.8	**.410**	.324	**.095**	.028	**.367**	**.062**
1.0	.354	.240	.070	.006	.322	.057
MB	.362	.246	.089	.003		
MC	.347	.350	.054	.059		

Table 3: Our evaluation scores for various values of λ. Smaller values of λ favor greater contrastiveness. Note that $\lambda = 1$ should be considered a baseline, because at this value the algorithm ignores the contrastiveness and it becomes a standard summarization problem. MB and MC refer to our model-based baselines described in Subsection 4.3.2. Bold scores are significant over all baselines according to a paired t-test.

micro-level evaluation due to overlapping pairs in the evaluation set. In this case, the ROUGE score is defined as the maximum score among all possible reference summaries (Lin, 2004).

We measure both unigram (removing stop words, denoted S-1) and bigram (retaining stop words, denoted S-2) recall, stemming words in all cases.

4.3.4 Evaluation Results

In order to evaluate our Comparative LexRank algorithm by itself, in this subsection we will not use the output of TAM as part of our summarization input, and will assign excerpts fixed values of $P(v|x) = 1$ for the correct label and 0 otherwise. We constructed our sentence vectors with unigrams (removing stop words) and no IDF weighting.

We set the PageRank damping factor (Erkan and Radev, 2004) to 0.01 and tried combinations of the redundancy threshold $\delta \in \{0.01, 0.05, 0.1, 0.2\}$ with different values of λ, the parameter which controls the level of contrastiveness. For each value of λ, we optimized δ on the original data set according to $S_{rep} \times S_{opp}$ so that we can directly compare these scores, and then we tuned δ separately for S_p. The summary length is 6 excerpts. To obtain more robust results, we repeated the experiment 100 times on random half-size subsets of our data. The scores shown in Table 3 are averaged across these trials.

In general, increasing λ increases S_{rep}, which suggests that tuning λ behaves as expected, and high- and mid-range λ values indeed produce summaries where the summaries of the two viewpoints are more similar to each other. Similarly, mid-range λ values produce substantially higher values of S_p-1, the unigram ROUGE scores for the micro contrastive summary, although there is not a large difference between the bigram scores. An example of our micro-level output is shown in Table 4.

As for our model-based baseline, we show results for both the basic algorithm (denoted MB) in addition to the contrastive modification (denoted MC). We see that the contrastive modification behaves as expected and produces much higher scores for S_{opp}, however, this method does not outperform our LexRank algorithm. It is interesting to note that in almost all cases where a contrastive objective is introduced, the scores for the opposite viewpoint S_{opp} increase without decreasing the S_{rep} scores, suggesting that contrastiveness can be introduced into a multi-view summarization problem without diminishing the overall quality of the summary. It is admittedly difficult to make generalizations about these methods from experiments with only one data set, but we have at least some evidence that our algorithm works as intended.

4.4 Unsupervised Summarization

So far we have focused on evaluating our viewpoint clustering models and our multi-view summarization algorithms separately. We will finally show how these two stages might work in tandem in unsupervised summarization of the Bitterlemons corpus.

Without a gold set, it is difficult to perform an extensive automatic evaluation as we did with the healthcare data. Instead we will perform a simple qualitative evaluation to see if the algorithm appears to achieve its goal. Thus, we asked 8 people to guess if each viewpoint's summary was written by Israeli or Palestinian authors. To diversify the summaries, for each annotator we randomly split each summary into two equal-sized subsets of the sentence set. Thus each person was asked to label four different summaries, which were presented in a random order. If humans can correctly identify the viewpoints, then this would suggest both that the TAM accurately clustered documents by viewpoint and the summarization algorithm is selecting sentences that coherently represent the viewpoints.

We first ran TAM on our data using the same procedure and parameters as in Subsection 4.2 using the *full-tuple* features. We repeated this 10 times and used the model that gave the highest data likelihood as our model for summarization input. We then gen-

For the Healthcare Bill	*Against the Healthcare Bill*
the government already provides half of the healthcare dollars in the united states [...] [they] might as well spend their dollars smarter	government is too much involvement.
my kids are uninsured.	a lot of people will be getting it that should be getting it on their own, and my kids will be paying a lot of taxes.
so everybody would have it and afford it.	we cannot afford it.
because of my family.	i don't know enough about it and i don't know where exactly it's going to put my family.
because i have no health insurance and i need it.	because i have health insurance.
cost of healthcare is so high.	high costs.

Table 4: An example of our *micro-level* contrastive summarization output on the healthcare data, using $\delta = 0.05$ and $\lambda = 0.5$.

erated macro contrastive summaries of our data for the two viewpoints with 6 sentences per viewpoint. We used unigram sentence vectors with IDF weighting. We used $\lambda = 0.5$ and $\delta = 0.1$, which gave the highest score at this λ value on the healthcare data.

Only one of these sentences was clustered incorrectly by TAM. The human judges correctly labeled 78% of the summary sets, suggesting that our system accurately selected some sentences that could be recognized as belonging to the viewpoints, but is not perfect. Unsupervised micro-level summaries were less coherent. Many of the sentences are mislabeled, and the ones that are correctly labeled are not representative of the collection.

This is not surprising, and indeed exposes the challenge inherent in our problem definition: clustering documents based on similarity and then highlighting sentences with high similarity but opposite cluster membership are almost conflicting objectives for an unsupervised learner. Such contrastive pairs are perhaps the most difficult data points to model. A good test of a viewpoint model may be whether it can capture the nuanced properties of the viewpoints needed to contrast them at the micro level.

5 Discussion

The properties of the text which we attempt to summarize in our work are related to the concept of *framing* from political science (Chong and Druckman, 2010), which is defined as "an interpretation or evaluation of an issue, event, or person that emphasizes certain of its features or consequences" focusing on "certain features and implications of the issue – rather than others." For example, someone in favor of the healthcare bill might focus on the benefits and someone against the bill might focus on the cost.

However, our approach is different in that our contrastive objective encourages the summaries to include each point as addressed by all viewpoints, rather than each viewpoint selectively emphasizing only certain points. In a sense, this makes our summary more like a live debate, where one side must directly respond to a point raised by the other side. For example, someone in favor of healthcare reform might cite the high cost of the current system, but someone against this might counter-argue that the proposed system in the new bill has its own high costs (as seen in the last row of Table 4). The idea is to show how both sides address the same issues.

Thus, we can say that we are summarizing the key arguments/issues/points from different opinions. Futhermore, our models and algorithms are defined very generally, and while we tested their viability in the domain of political opinion, they may also be useful for many other comparative tasks.

In conclusion, we have presented steps toward a two-stage system that can automatically extract and summarize viewpoints in opinionated text. First, we have shown that accuracy of clustering documents by viewpoint can be enhanced by using simple but rich dependency features. This can be done within the framework of existing probabilistic topic models without altering the models simply by using a "bag of features" representation of documents.

Second, we have introduced *Comparative LexRank*, an extension of the LexRank algorithm that aims to generate contrastive summaries both at the macro and micro level. The algorithm presented is general enough that it can be applied to any number of viewpoints, and can accomodate input where the viewpoints are either given fixed labels, or given probabilistic assignments. The tradeoff between contrast and representation can flexibly be tuned to an application's needs.

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Samuel Brody and Noemie Elhadad. 2010. An unsupervised aspect-sentiment model for online reviews. In *NAACL '10*.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98*, pages 335–336.

Dennis Chong and James N. Druckman. 2010. Identifying frames in political news. In Erik P. Bucy and R. Lance Holbert, editors, *Sourcebook for Political Communication Research: Methods, Measures, and Analytical Techniques*. Routledge.

Cindy Chung and James W. Pennebaker. 2007. The psychological function of function words. *Social Communication: Frontiers of Social Psychology*, pages 343–359.

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, 22(1):457–479.

Stephan Greene and Philip Resnik. 2009. More than words: syntactic packaging and implicit sentiment. In *NAACL '09*, pages 503–511.

Aria Haghighi and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. In *NAACL '09*, pages 362–370.

Sanda Harabagiu, Andrew Hickl, and Finley Lacatusu. 2006. Negation, contrast and contradiction in text processing.

Minqing Hu and Bing Liu. 2004. Mining opinion features in customer reviews. In *Proceedings of AAAI*, pages 755–760.

Minqing Hu and Bing Liu. 2006. Opinion extraction and summarization on the Web. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-2006), Nectar Paper Track*, Boston, MA.

Jeffrey M. Jones. 2010. "in u.s., 45% favor, 48% oppose obama healthcare plan", March.

Mahesh Joshi and Carolyn Penstein Rosé. 2009. Generalizing dependency features for opinion mining. In *ACL-IJCNLP '09: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 313–316.

Hyun Duk Kim and ChengXiang Zhai. 2009. Generating comparative summaries of contradictory opinions in text. In *CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management*, pages 385–394, New York, NY, USA. ACM.

Kevin Lerman and Ryan McDonald. 2009. Contrastive summarization: an experiment with consumer reviews. In *NAACL '09*, pages 113–116, Morristown, NJ, USA. Association for Computational Linguistics.

Wei-Hao Lin, Theresa Wilson, Janyce Wiebe, and Alexander Hauptmann. 2006. Which side are you on?: identifying perspectives at the document and sentence levels. In *CoNLL-X '06: Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 109–116.

Wei-Hao Lin, Eric Xing, and Alexander Hauptmann. 2008. A joint topic and perspective model for ideological discourse. In *ECML PKDD '08: Proceedings of the European conference on Machine Learning and Knowledge Discovery in Databases - Part II*, pages 17–32, Berlin, Heidelberg. Springer-Verlag.

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

Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion observer: analyzing and comparing opinions on the web. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pages 342–351, New York, NY, USA. ACM Press.

Marie-Catherine De Marneffe and Christopher Manning. 2008. Stanford typed dependencies manual. Technical report, Stanford University.

Marie-Catherine De Marneffe, Anna Rafferty, and Christopher Manning. 2008. Finding contradictions in text. In *Proceedings of the Association for Computational Linguistics Conference (ACL)*.

Ryan McDonald. 2007. A study of global inference algorithms in multi-document summarization. In *ECIR'07: Proceedings of the 29th European conference on IR research*, pages 557–564, Berlin, Heidelberg. Springer-Verlag.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1998. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project.

Michael Paul and Roxana Girju. 2010. A two-dimensional topic-aspect model for discovering multi-faceted topics. In *AAAI-2010: Twenty-Fourth Conference on Artificial Intelligence*.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354.

Li Zhuang, Feng Jing, Xiao-yan Zhu, and Lei Zhang. 2006. Movie review mining and summarization. In *Proceedings of the ACM SIGIR Conference on Information and Knowledge Management (CIKM)*.