Contrastive Summarization: An Experiment with Consumer Reviews

Kevin Lerman
Columbia University
New York, NY
klerman@cs.columbia.edu

Ryan McDonald
Google Inc.
New York, NY
ryanmcd@google.com

Abstract

Contrastive summarization is the problem of jointly generating summaries for two entities in order to highlight their differences. In this paper we present an investigation into contrastive summarization through an implementation and evaluation of a contrastive opinion summarizer in the consumer reviews domain.

1 Introduction

Automatic summarization has historically focused on summarizing events, a task embodied in the series of Document Understanding Conferences[1]. However, there has also been work on *entity-centric summarization*, which aims to produce summaries from text collections that are relevant to a particular entity of interest, e.g., product, person, company, etc. A well-known example of this is from the opinion mining community where there has been a number of studies on summarizing the expressed sentiment towards entities (cf. Hu and Liu (2006)). Another recent example of entity-centric summarization is the work of Filippova et al. (2009) to produce company-specific financial report summaries.

In this study we investigate a variation of entity-centric summarization where the goal is not to summarize information about a single entity, but pairs of entities. Specifically, our aim is to jointly generate two summaries that highlight differences between the entities – a task we call *contrastive summarization*. An obvious application comes from the consumer reviews domain, where a person considering a purchase wishes to see the differences in opinion about the top candidates without reading all the reviews for each product. Other applications include contrasting financial news about related companies or comparing platforms of political candidates.

Contrastive summarization has many points of comparison in the NLP, IR and Data-Mining literature. Jindal and Liu (2006) introduce techniques to find and analyze explicit comparison sentences, but this assumes that such sentences exist. In contrastive summarization, there is no assumption that two entities have been explicitly compared. The goal is to automatically generate the comparisons based on the data. In the IR community, Sun et al. (2006) explores retrieval systems that align query results to highlight points of commonality and difference. In contrast, we attempt to identify contrasts from the data, and then generate summaries that highlight them. The *novelty detection task* of determining whether a new text in a collection contains information distinct from that already gathered is also related (Soboroff and Harman, 2005). The primary difference here is that contrastive summarization aims to extract information from one collection not present in the other in addition to information present in both collections that highlights a difference between the entities.

This paper describes a contrastive summarization experiment where the goal is to generate contrasting opinion summaries of two products based on consumer reviews of each. We look at model design choices, describe an implementation of a contrastive summarizer, and provide an evaluation demonstrating a significant improvement in the usefulness of contrastive summaries versus summaries generated by single-product opinion summarizers.

2 Single-Product Opinion Summarization

As input we assume a set of relevant text excerpts (typically sentences), $T = \{t_1, \ldots, t_m\}$, which con-

[1] http://duc.nist.gov/

Proceedings of NAACL HLT 2009: Short Papers, pages 113–116,
Boulder, Colorado, June 2009.

tain opinions about some product of interest. The goal of opinion summarization[2] is to select some number of text excerpts to form a summary S of the product so that S is representative of the average opinion and speaks to its important aspects (also proportional to opinion), which we can formalize as:

$$S = \arg\max_{S \subseteq T} \mathcal{L}(S) \quad \text{s.t. } \textsc{length}(S) \leq K$$

where $\mathcal{L}$ is some score over possible summaries that embodies what a user might desire in an opinion summary, $\textsc{length}(S)$ is the length of the summary and K is a pre-specified length constraint.

We assume the existence of standard sentiment analysis tools to provide the information used in the scoring function $\mathcal{L}$. First, we assume the tools can assign a sentiment score from -1 (negative) to 1 (positive) to an arbitrary span of text. Second, we assume that we can extract a set of aspects that the text is discussing (e.g, "The sound was crystal clear" is about the aspect *sound quality*). We refer the reader to abundance of literature on sentiment analysis for more details on how such tools can be constructed (cf. Pang and Lee (2008)). For this study, we use the tools described and evaluated in Lerman et al. (2009). We note however, that the subject of this discussion is not the tools themselves, but their use.

The single product opinion summarizer we consider is the Sentiment Aspect Match model (SAM) described and evaluated in (Lerman et al., 2009). Underlying SAM is the assumption that opinions can be described by a bag-of-aspects generative process where each aspect is generated independently and the sentiment associated with the aspect is generated conditioned on its identity,

$$p(t) = \prod_{a \in A_t} p(a) p(\textsc{sent}(a_t)|a)$$

where A_t is a set of aspects that are mentioned in text excerpt t, $p(a)$ is the probability of seeing aspect a, and $\textsc{sent}(a_t) \in [-1, 1]$ is the sentiment associated with aspect a in t. The SAM model sets $p(a)$ through the maximum likelihood estimates over T and assumes $p(\textsc{sent}(a_t)|a)$ is normally distributed with a mean and variance also estimated from T. We denote $\textsc{sam}(T)$ as the model learned using the entire set of candidate text excerpts T.

The SAM summarizer scores each potential summary, S, by learning another model $\textsc{sam}(S)$ based on the text excerpts used to construct S. We can then measure the distance between a model learned over the full set T and a summary S by summing the KL-divergence between their learned probability distributions. In our case we have $1 + |A_T|$ distributions – $p(a)$, and $p(\cdot|a)$ for all $a \in A_T$. We then define $\mathcal{L}$:

$$\mathcal{L}(S) = -\textsc{kl}(\textsc{sam}(T), \textsc{sam}(S))$$

That is, the SAM summarizer prefers summaries whose induced model is close to the model induced for all the opinions about the product of interest. Thus, a good summary should (1) mention aspects in roughly the same proportion that they are mentioned in the full set of opinions *and* (2) mention aspects with sentiment also in proportion to what is observed in the full opinion set. A high scoring summary is found by initializing a summary with random sentences and hill-climbing by replacing sentences one at a time until convergence.

We chose to use the SAM model for our experiment for two reasons. First, Lerman et al. (2009) showed that among a set of different opinion summarizers, SAM was rated highest in a user study. Secondly, as we will show in the next section, the SAM summarization model can be naturally extended to produce contrastive summaries.

3 Constrastive Summarization

When jointly generating pairs of summaries, we attempt to highlight differences between two products. These differences can take multiple forms. Clearly, two products can have different prevailing sentiment scores with respect to an aspect (e.g. "Product X has great image quality" vs "Product Y's image quality is terrible"). Reviews of different products can also emphasize different aspects. Perhaps one product's screen is particularly good or bad, but another's is not particularly noteworthy – or perhaps the other product simply doesn't have a screen. Regardless of sentiment, reviews of the first product will emphasize the screen quality aspect more than those of the second, indicating that our summary should as well.

[2] We focus on text-only opinion summaries as opposed to those based on numeric ratings (Hu and Liu, 2006).

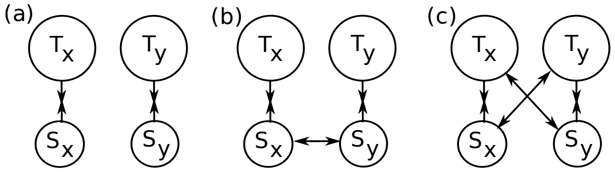

Figure 1: (a) Non-joint model: Generates summaries for two products independently. (b) Joint model: Summaries attempt to look like text they are drawn from, but contrast each-other. (c) Joint model: Like (b), except summaries contrast text that the other summary is drawn from.

As input to our contrastive summarizer we assume two products, call them x and y as well as two corresponding candidate sets of opinions, T_x and T_y, respectively. As output, a contrastive summarizer will produce two summaries – S_x for product x and S_y for product y – so that the summaries highlight the differences in opinion between the two products.

What might a contrastive summarizer look like on a high-level? Figure 1 presents some options. The first example (1a) shows a system where each summary is generated independently, i.e., running the SAM model on each product separately without regard to the other. This procedure may provide some useful contrastive information, but any such information will be present incidentally. To make the summaries specifically contrast each other, we can modify our system by explicitly modeling the fact that we want summaries S_x and S_y to contrast. In the SAM model this is trivial as we can simply add a term to the scoring function $\mathcal{L}$ that attempts to maximize the KL-divergence between the two summaries induced models $\textsc{sam}(S_x)$ and $\textsc{sam}(S_y)$.

This approach is graphically depicted in figure 1b, where the system attempts to produce summaries that are maximally similar to the opinion set they are drawn from and minimally similar from each other. However, some obvious degenerate solutions arise if we chose to model our system this way. Consider two products, x and y, for which all opinions discuss two aspects a and b with identical frequency and sentiment polarity. Furthermore, several opinions of x and y discuss an aspect c, but with opposite sentiment polarity. Suppose we have to build contrastive summaries and only have enough space to cover a single aspect. The highest scoring contrastive pair of summaries would consist of one for x that mentions a exclusively, and one for y that mentions b exclusively – these summaries each mention a promiment aspect of their product, and have no overlap with each other. However, they provide a false contrast because they each attempt to contrast the other *s*ummary, rather than the other *p*roduct. Better would be for both to cover aspect c.

To remedy this, we reward summaries that instead have a high KL-divergence with respect to the other product's *full* model $\textsc{sam}(T)$ as depicted in Figure 1c. Under this setup, the degenerate solution described above is no longer appealing, as both summaries have the same KL-divergence with respect to the other product as they do to their own product. The fact that the summaries themselves are dissimilar is irrelevant. Comparing the summaries only to the products' full language models prevents us from rewarding summaries that convey a false contrast between the products under comparison. Specifically, we now optimize the following joint summary score:

$$\begin{aligned}\mathcal{L}(S_x, S_y) \quad = \quad & -\textsc{kl}(\textsc{sam}(T_x), \textsc{sam}(S_x)) \\ & -\textsc{kl}(\textsc{sam}(T_y), \textsc{sam}(S_y)) \\ & +\textsc{kl}(\textsc{sam}(T_x), \textsc{sam}(S_y)) \\ & +\textsc{kl}(\textsc{sam}(T_y), \textsc{sam}(S_x))\end{aligned}$$

Note that we could additionally model divergence between the two summaries (i.e., merging models in figures 1b and c), but such modeling is redundant. Furthermore, by not explicitly modeling divergence between the two summaries we simplify the search space as each summary can be constructed without knowledge of the content of the second summary.

4 The Experiment

Our experiments focused on consumer electronics. In this setting an entity to be summarized is one specific product and T is a set of segmented user reviews about that product. We gathered reviews for 56 electronics products from several sources such as CNet, Epinions, and PriceGrabber. The products covered 15 categories of electronics products, including MP3 players, digital cameras, laptops, GPS systems, and more. Each had at least four reviews, and the mean number of reviews per product was 70.

We manually grouped the products into categories (MP3 players, cameras, printers, GPS sys-

System	As Received	Consolidated
SAM	1.85 ± 0.05	1.82 ± 0.05
SAM + contrastive	**1.76** ± 0.05	**1.68** ± 0.05

Table 1: Mean rater scores for contrastive summaries by system. Scores range from 0-3 and lower is better.

tems, headphones, computers, and others), and generated contrastive summaries for each pair of products in the same category using 2 different algorithms: (1) The SAM algorithm for each product individually (figure 1a) and (2) The SAM algorithm with our adaptation for contrastive summarization (figure 1c). Summaries were generated using $K = 650$, which typically consisted of 4 text excerpts of roughly 160 characters. This allowed us to compare different summaries without worrying about the effects of summary length on the ratings. In all, we gathered 178 contrastive summaries (89 per system) to be evaluated by raters and each summary was evaluated by 3 random raters resulting in 534 ratings. The raters were 55 everyday internet users that signed-up for the experiment and were assigned roughly 10 random ratings each. Raters were shown two products and their contrastive summaries, and were asked to list 1-3 differences between the products as seen in the two summaries. They were also asked to read the products' reviews to help ensure that the differences observed were not simply artifacts of the summarizer but in fact are reflected in actual opinions. Finally, raters were asked to rate the helpfulness of the summaries in identifying these distinctions, rating each with an integer score from 0 (”extremely useful”) to 3 (”not useful”).

Upon examining the results, we found that raters had a hard time finding a meaningful distinction between the two middle ratings of 1 and 2 (“useful” and “somewhat useful”). We therefore present two sets of results: one with the scores as received from raters, and another with all 1 and 2 votes consolidated into a single class of votes with numerical score 1.5. Table 1 gives the average scores per system, lower scores indicating superior performance.

5 Analysis and Conclusions

System	2+ raters	All 3 raters
SAM	0.8	0.2
SAM + contrastive	**2.0**	**0.6**

Table 2: Average number of points of contrast per comparison observed by multiple raters, by system. Raters were asked to list up to 3. Higher is better.

The scores indicate that the addition of the contrastive term to the SAM model improves helpfulness, however both models roughly have average scores in the somewhat-useful to useful range. The difference becomes more pronounced when looking at the consolidated scores. The natural question arises: does the relatively small increase in helpfulness reflect that the contrastive summarizer is doing a poor job? Or does it indicate that users only find slightly more utility in contrastive information in this domain? We inspected comments left by raters in an attempt to answer this. Roughly 80% of raters were able to find at least two points of contrast in summaries generated by the SAM+contrastive versus 40% for summaries generated by the simple SAM model. We then examined the consistency of rater comments, i.e., to what degree did different raters identify the same points of contrast from a specific comparison? We report the results in table 2. Note that by this metric in particular, the contrastive summarizer outperforms its the single-product summarizer by significant margins and provides a strong argument that the contrastive model is doing its job.

Acknowledgements: The Google sentiment analysis team for insightful discussions and suggestions.

References

K. Filippova, M. Surdeanu, M. Ciaramita, and H. Zaragoza. 2009. Company-oriented extractive summarization of financial news. In *Proc. EACL*.

M. Hu and B. Liu. 2006. Opinion extraction and summarization on the web. In *Proc. AAAI*.

N. Jindal and B. Liu. 2006. Mining comparative sentences and relations. In *Proc. AAAI*.

Kevin Lerman, Sasha Blair-Goldensohn, and Ryan McDonald. 2009. Sentiment summarization: Evaluating and learning user preferences. In *Proc. EACL*.

B. Pang and L. Lee. 2008. *Opinion mining and sentiment analysis*. Now Publishers.

I. Soboroff and D. Harman. 2005. Novelty detection: The TREC experience. In *Proc. HLT/EMNLP*.

Sun, Wang, Shen, Zeng, and Chen. 2006. CWS: A Comparative Web search System. In *Proc. WWW*.