Empirical analysis of exploiting review helpfulness for extractive summarization of online reviews

Wenting Xiong
University of Pittsburgh
Department of Computer Science
wex12@cs.pitt.edu

Diane Litman
University of Pittsburgh
Department of Computer Science & LRDC
litman@cs.pitt.edu

Abstract

We propose a novel unsupervised extractive approach for summarizing online reviews by exploiting review helpfulness ratings. In addition to using the helpfulness ratings for review-level filtering, we suggest using them as the supervision of a topic model for sentence-level content scoring. The proposed method is metadata-driven, requiring no human annotation, and generalizable to different kinds of online reviews. Our experiment based on a widely used multi-document summarization framework shows that our helpfulness-guided review summarizers significantly outperform a traditional content-based summarizer in both human evaluation and automated evaluation.

1 Introduction

Multi-document summarization has great potential in online reviews, as manually reading comments provided by other users is time consuming if not impossible. While extractive techniques are generally preferred over abstractive ones (as abstraction can introduce disfluency), existing extractive summarizers are either supervised or based on heuristics of certain desired characteristics of the summarization result (e.g., maximize n-gram coverage (Nenkova and Vanderwende, 2005), etc.). However, when it comes to online reviews, there are problems with both approaches: the first one requires manual annotation and is thus less generalizable; the second one might not capture the salient information in reviews from different *domains* (*camera* reviews vs. *movie* reviews), because the heuristics are designed for traditional genres (e.g., news articles) while the utility of reviews might vary with the review domain.

We propose to exploit review metadata, that is *review helpfulness ratings*[1], to facilitate review summarization. Because this is user-provided feedback on review helpfulness which naturally reflects users' interest in online review exploration, our approach captures domain-dependent salient information adaptively. Furthermore, as this metadata is widely available online (e.g., Amazon.com, IMDB.com)[2], our approach is unsupervised in the sense that no manual annotation is needed for summarization purposes. Therefore, we hypothesize that summarizers guided by review helpfulness will outperform systems based on textual features/heuristics designed for traditional genres. To build such helpfulness-guided summarizers, we introduce review helpfulness during content selection in two ways: 1) using the review-level helpfulness ratings directly to filter out unhelpful reviews, 2) using sentence-level helpfulness features derived from review-level helpfulness ratings for sentence scoring. As we observe in our pilot study that supervised LDA (sLDA) (Blei and McAuliffe, 2010) trained with review helpfulness ratings has potential in differentiating review helpfulness at the sentence level, we develop features based on the inferred hidden topics from sLDA to capture the helpfulness of a review sentence for summarization purposes. We implement our helpfulness-guided review summarizers based on an widely used open-source multi-document extractive summarization framework (MEAD (Radev et al., 2004)). Both human and

[1]This is the percentage of readers who found the review to be helpful (Kim et al., 2006).

[2]If it is not available, the review helpfulness can be assessed fully automatically (Kim et al., 2006; Liu et al., 2008).

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 1985–1995, Dublin, Ireland, August 23-29 2014.

automated evaluations show that our helpfulness-guided summarizers outperform a strong baseline that MEAD provides across multiple review domains. Further analysis on the human summaries shows that some effective heuristics proposed for traditional genres might not work well for online reviews, which indirectly supports our use of review metadata as supervision. The presented work also extrinsically demonstrates that the helpfulness-related topics learned from the review-level supervision can capture review helpfulness at the sentence-level.

2 Related Work

In multi-document extractive summarization, various unsupervised approaches have been proposed to avoid manual annotation. A key task in extractive summarization is to identify important text units. Prior successful extractive summarizers score a sentence based on n-grams within the sentence: by the word frequency (Nenkova and Vanderwende, 2005), bigram coverage (Gillick and Favre, 2009), topic signatures (Lin and Hovy, 2000) or latent topic distribution of the sentence (Haghighi and Vanderwende, 2009), which all aim to capture the "core" content of the text input. Other approaches regard the n-gram distribution difference (e.g., Kullback-Lieber (KL) divergence) between the input documents and the summary (Lin et al., 2006), or based on a graph-representation of the document content (Erkan and Radev, 2004; Leskovec13 et al., 2005), with an implicit goal to maximize the output representativeness. In comparison, while our approach follows the same extractive summarization paradigm, it is metadata driven, identifying important text units through the guidance of user-provided review helpfulness assessment.

When it comes to online reviews, the desired characteristics of a review summary are different from traditional text genres (e.g., news articles), and could vary from one review domain to another. Thus different review summarizers have been proposed to focus on different desired properties of review summaries, primarily based on opinion mining and sentiment analysis (Carenini et al., 2006; Lerman et al., 2009; Lerman and McDonald, 2009; Kim and Zhai, 2009). Here the desired property varies from the coverage of product aspects (Carenini et al., 2006; Lerman et al., 2009) to the degree of agreement on aspect-specific sentiment (Lerman et al., 2009; Lerman and McDonald, 2009; Kim and Zhai, 2009). While there is a large overlap between text summarization and review opinion mining, most work focuses on sentiment-oriented aspect extraction and the output is usually a set of topics words plus their representative text units (Hu and Liu, 2004; Zhuang et al., 2006). However, such a topic-based summarization framework is beyond the focus of our work, as we aim to adapt traditional extractive techniques to the review domain by introducing review helpfulness ratings as guidance.

In this paper, we utilize review helpfulness via using sLDA. The idea of using sLDA in text summarization is not new. However, the model is previously applied at the sentence level (Li and Li, 2012), which requires human annotation on the sentence importance. In comparison, our use of sLDA is at the document (review) level, using existing metadata of the document (review helpfulness ratings) as the supervision, and thus requiring no annotation at all. With respect to the use of review helpfulness ratings, early work of review summarization (Liu et al., 2007) only consider it as a filtering criteria during input preprocessing. Other researchers use it as the gold-standard for automated review helpfulness prediction, a predictor of product sales (Ghose and Ipeirotis, 2011), a measurement of reviewers' authority in social network analysis (Lu et al., 2010), etc.

3 Helpfulness features for sentence scoring

While the most straightforward way to utilize review helpfulness for content selection is through filtering (Liu et al., 2007) (further discussed in Section 4.3), we also propose to take into account review helpfulness during sentence scoring by learning helpfulness-related review topics in advance. Because sLDA learns the utility of the topics for predicting review-level helpfulness ratings (decomposing review helpfulness ratings by topics), we develop novel features (*rHelpSum* and *sHelpSum*) based on the inferred topics of the words in a sentence to capture its helpfulness in various perspectives. We later use them for sentence scoring in a helpfulness-guided summarizer (Section 4.3).

Compared with LDA (Blei et al., 2003), sLDA (Blei and McAuliffe, 2010) introduces a response

variable $y_i \in Y$ to each document D_i during topic discovery. The model not only learns the topic assignment $z_{1:N}$ for words $w_{1:N}$ in D_i, it also learns a function from the posterior distribution of z in D to Y. When Y is the review-level helpfulness gold-standard, the model leans a set of topics predictive of review helpfulness, as well as the utility of z in predicting review helpfulness y_i, denoted as η. (Both z and η are K-dimensional.)

At each inference step, sLDA assigns a topic ID to each word in every review. $z_l = k$ means that the topic ID for word at position l in sentence s is k. Given the topic assignments $z_{1:L}$ to words $w_{1:L}$ in a review sentence s, we estimate the contribution of s to the helpfulness of the review it belongs to (Formula 1), as well as the average topic importance in s (Formula 2). While $rHelpSum$ is sensitive to the review length, $sHelpSum$ is sensitive to the sentence length.

$$rHelpSum(s) = \frac{1}{N} \sum_{l=1}^{l=L} \sum_{k} \eta_k p(z_l = k) \tag{1}$$

$$sHelpSum(s) = \frac{1}{L} \sum_{l=1}^{l=L} \sum_{k} \eta_k p(z_l = k) \tag{2}$$

As the topic assignment in each inference iteration might not be the same, Riedl and Biemann (Riedl and Biemann, 2012) proposed the *mode* method in their application of LDA for text segmentation – use the most frequently assigned topic for each word in all iterations as the final topic assignment – to address the instability issue. Inspired by their idea, we also use the *mode* method to infer the topic assignment in our task, but only apply the *mode* method to the last 10 iterations, because the topic distribution might not be well learned at the beginning.

4 Experimental setup

To investigate the utility of exploiting user-provided review helpfulness ratings for content selection in extractive summarization, we develop two helpfulness-guided summarizers based on the MEAD framework (HelpfulFilter and HelpfulSum). We compare our systems' performance against a strong unsupervised extractive summarizer that MEAD supports as our baseline (MEAD+LexRank). To focus on sentence scoring only, we use the same MEAD word-based MMR (Maximal Marginal Relevance) reranker (Carbonell and Goldstein, 1998) for all summarizers, and set the length of the output to be 200 words.

4.1 Data

Our data consists of two kinds of online reviews: 4050 Amazon camera reviews provided by Jindal and Liu (2008) and 280 IMDB movie reviews that we collected by ourselves. Both corpora were used in our prior work of automatically predicting review helpfulness, in which every review has at least three helpfulness votes. On average, the helpfulness of camera reviews is .80 and that of movie reviews is .74.

Summarization test sets. Because the proposed approach method is purely unsupervised, and we do not optimize our summarization parameters during learning, we evaluate our approach based on a subset of review items directly: we randomly sample 18 reviews for each review item (a camera or movie) and randomly select 3 items for each review domain. In total there are 6 summarization test sets (3 items $\times$ 2 domains), where each contains 18 reviews to be summarized (i.e. "summarizing 18 camera reviews for Nikon D3200"). In the summarization test sets, the average number of sentences per review is 9 for camera reviews, and 18 for movie reviews; the average number of words per sentence in the camera reviews and movie reviews are 25 and 27, respectively.

4.2 sLDA training

We implement sLDA based on the topic modeling framework of Mallet (McCallum, 2002) using 20 topics ($K = 20$) and the best hyper-parameters (topic distribution priors α and word distribution priors

β) that we learned in our pilot study on LDA. [3]

Since our summarization approach is unsupervised, we learn the topic assignment for each review word using the corresponding sLDA model trained on all reviews of that domain (4050 reviews for camera and 280 reviews for movie).[4]

4.3 Three summarizers

Baseline (MEAD+LexRank): The default feature set of MEAD includes *Position*, *Length*, and *Centroid*. Here *Length* is a word-count threshold, which gives score 0 to sentences shorter than the threshold. As we observe that short review sentences sometimes can be very informative as well (e.g., "This camera is so amazing!", "The best film I have ever seen!"), we adjust $Length$ to 5 from its default value 9. MEAD also provides scripts to compute *LexRank* (Erkan and Radev, 2004), which is a more advanced feature using graph-based algorithm for computing relative importance of textual units. We supplement the default feature set with *LexRank* to get the best summarizer from MEAD, yielding the sentence scoring function $F_{baseline}(s)$, in which s is a given sentence and all features are assigned equal weights (same as in the other two summarizers).

$$F_{baseline}(s) = \begin{cases} Position + Centroid + LexRank & \text{if } Length \geq 5 \\ 0 & \text{if } Length < 5 \end{cases} \quad (3)$$

HelpfulFilter: This summarizer is a direct extension of the baseline, which considers review-level helpfulness ratings ($hRating$) as an additional filtering criteria in its sentence scoring function $F_{HelpfulFilter}$. (In our study, we omit the automated prediction (Kim et al., 2006; Liu et al., 2008) and filter reviews by their helpfulness gold-standard directly.) We set the cutting threshold to be the average helpfulness rating of all the reviews that we used to train the topic model for the corresponding domain ($hRatingAve(domain)$).

$$F_{HelpfulFilter}(s) = \begin{cases} F_{baseline}(s) & \text{if } hRating(s) \geq hRatingAve(domain) \\ 0 & \text{if } hRating(s) < hRatingAve(domain) \end{cases} \quad (4)$$

HelpfulSum: To isolate the contribution of review helpfulness, the second summarizer only uses helpfulness related features in its sentence scoring function $F_{HelpfulSum}$. The features are ***rHelpSum*** – the contribution of a sentence to the overall helpfulness of its corresponding review, ***sHelpSum*** – the average topic weight in a sentence for predicting the overall helpfulness of the review (Formula 1 and 2), plus $hRating$ for filtering. Note that there is no overlap between features used in the baseline and HelpfulSum, as we wonder if the helpfulness information alone is good enough for discovering salient review sentences.

$$F_{HelpfulSum}(s) = \begin{cases} rHelpSum(s) + sHelpSum(s) & \text{if } hRating(s) \geq hRatingAve(domain) \\ 0 & \text{if } hRating(s) < hRatingAve(domain) \end{cases} \quad (5)$$

5 Evaluation

For evaluation, we will first present our human evaluation user study and then present the automated evaluation result based on human summaries collected from the user study.

[3] In our pilot study, we experimented with various hyper-parameter settings, and trained the model with 100 sampling iterations in both the **E**stimation and the **M**aximization steps. As we found the best results are more likely to be achieved when $\alpha = 0.5$, $\beta = 0.1$, we use this setting to train the sLDA model in our summarization experiment.

[4] In practice, this means that we need to (re)train the topic model after given the summarization test set.

5.1 Human evaluation

The goal of our human evaluation is to compare the effectiveness of 1) using a traditional content selection method (MEAD+LexRank), 2) using the traditional method enhanced by review-level helpfulness filtering (HelpfulFilter), and 3) using sentence helpfulness features estimated by sLDA plus review-level helpfulness filtering (HelpfulSum) for building an extractive multi-document summarization system for online reviews. Therefore, we use a within-subject design in our user study for each review domain, considering the *summarizer* as the main effect on human evaluation results.

The user study is carried out in the form of online surveys (one survey per domain) hosted by Quadrics. In total, 36 valid users participated in our online-surveys.[5] We randomly assigned 18 of them to the camera reviews, and the rest 18 to the movie reviews.

5.1.1 Experimental procedures

Each online survey contains three summarization sets. The human evaluation on each one is taken in three steps:

Step 1: We first require users to perform **manual summarization**, by selecting 10 sentences from the input reviews (displayed in random order for each visit). This ensures that users are familiar with the input text so that they can have fair judgement on machine-generated results. To help users select the sentences, we provide an introductory scenario at the beginning of the survey to illustrate the potential application in accordance with the domain (e.g., Figure 1).

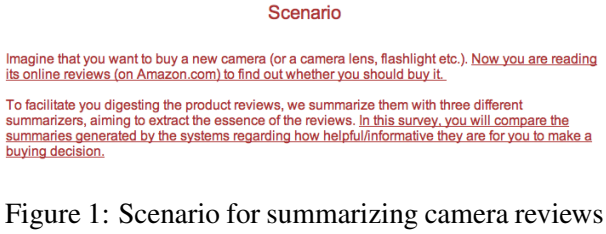

Figure 1: Scenario for summarizing camera reviews

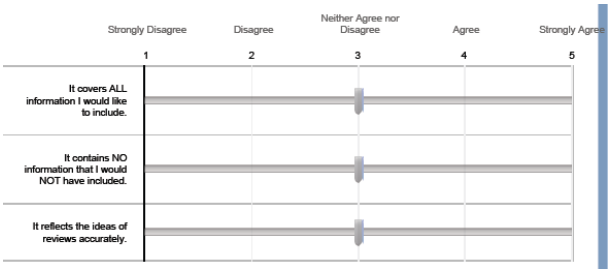

Figure 2: Content evaluation

Step 2: We then ask users to perform **pairwise comparison** on summaries generated by the three systems. The three pairs are generated in random order; and the left-or-right display position (in Figure 3) of the two summaries in each pair is also randomly selected. Here we use the same 5-level preference ratings used in (Lerman et al., 2009), and translate them into integers from -2 to 2 in our result analysis.
Step 3: Finally, we ask users to evaluate the three summaries in isolation regarding the summary quality in three content-related aspects: *recall*, *precision* and *accuracy* (top, middle and bottom in Figure 2, respectively), which were used in (Carenini et al., 2006). In this **content evaluation**, the three summaries are randomly visited and the users rate the proposed statements (one for each aspect) on a 5-point scale.

5.1.2 Results

Pairwise comparison. We use a mixed linear model to analyze user preference over the three summary pairs separately, in which "summarizer" is a between-subject factor, "review item" is the repeated factor, and "user" is a random effect. Results are summarized in Table 1. (Positive preference ratings on "A over B" means A is preferred over B; negative ratings means B is preferred over A.) As we can see, **HelpfulSum** is the best: it is consistently preferred over the other two summarizers across domains and the preference is significant throughout conditions except when compared with HelpfulFilter on movie reviews. **HelpfulFilter** is significantly preferred over the baseline (MEAD+LexRank) for movie reviews, while it does not outperform the baseline on camera reviews. A further look at the compression rate (cRate) of the three systems (Table 2) shows that on average HelpfulFilter generates shortest summaries

[5] All participants are older than eighteen, recruited via university mailing lists, on-campus flyers as well as social networks online. While we also considered educational peer reviews as a third domain, about half of the participants dropped out in the middle of the survey. Thus we only consider the two e-commerce domains in this paper.

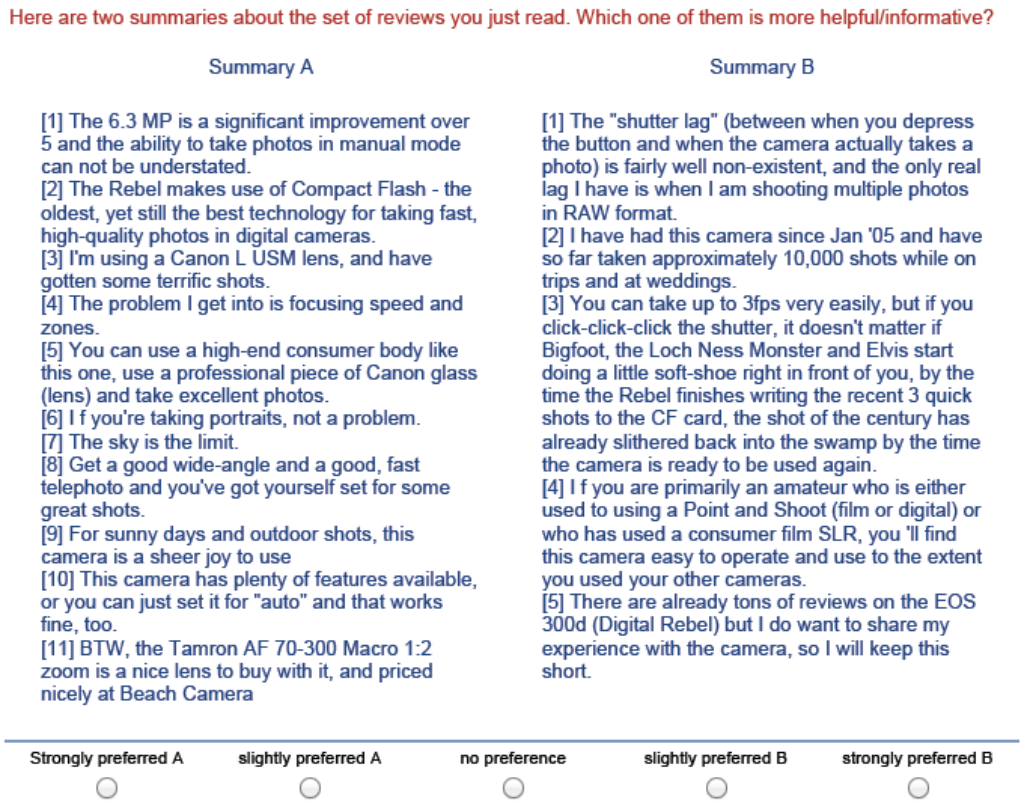

Figure 3: Example of pairwise comparison for summarizing camera reviews (left:HelpfulSum, right: the baseline).

among the three summarizers on camera reviews[6], which makes it naturally harder for **HelpfulFilter** to beat the other two (Napoles et al., 2011).

Pair	Domain	Est. Mean	Sig.
HelpfulFilter	Camera	-.602	.001
over MEAD+LexRank	Movie	.621	.000
HelpfulSum	Camera	.424	.011
over MEAD+LexRank	Movie	.601	.000
HelpfulSum	Camera	1.18	.000
over HelpfulFilter	Movie	.160	.310

Table 1: Mixed-model analysis of user preference ratings in pairwise comparison across domains. Confidence interval = 95%. The preference rating is ranged from -2 to 2.

Summarizer	Camera	Movie
MEAD+LexRank	6.07%	2.64%
HelpfulFilter	3.25%	2.39%
HelpfulSum	5.94%	2.69%
Human (Ave.)	6.11%	2.94%

Table 2: Compression rate of the three systems across domains.

Content evaluation. We summarize the average quality ratings (Figure 2) received by each summarizer across review items and users for each review domain in Table 3. We carry out paired T-tests for every pair of summarizers on each quality metric. While no significant difference is found among the three summarizers on any quality metric for movie reviews, there are differences for camera reviews. In terms of both accuracy and recall, HelpfulSum is significantly better than HelpfulFilter (p=.008 for accuracy, p=.034 for recall) and the baseline is significantly better than HelpfulFilter (p=.005 for accuracy, p=.005 for recall), but there is no difference between HelpfulSum and the baseline. For precision, no significant

[6]While we limit the summarization output to be 200 words in MEAD, as the content selection is at the sentence level, the summaries can have different number of words in practice. Considering that word-based MMR controls the redundancy in the selected summary sentences ($\lambda = 0.5$ as suggested), there might be enough content to select using $F_{HelpfulFilter}$.

difference is observed in either domain.

Summarizer	Camera			Movie		
Metric	Precision	Recall	Accuracy	Precision	Recall	Accuracy
MEAD+LexRank	2.63	**3.24**	3.57	2.50	2.59	2.93
HelpfulFilter	**2.78**	2.74	3.11	2.44	2.61	2.96
HelpfulSum	2.41	3.19	**3.69**	**2.52**	**2.67**	**3.02**

Table 3: Human ratings for content evaluation. The best result on each metric is bolded for every review domain (the higher the better).

With respect to pairwise evaluation, content evaluation yields consistent results on camera reviews between HelpfulFilter vs. the baseline and HelpfulSum vs. HelpfulFilter. However, only pairwise comparison (preference ratings) shows significant difference between HelpfulSum vs. the baseline and the difference in the summarizers' performance on movie reviews. This confirms that pairwise comparison is more suitable than content evaluation for human evaluation (Lerman et al., 2009).

5.2 Automated evaluation based on ROUGE metrics

Although human evaluation is generally preferred over automated metrics for summarization evaluation, we report our automated evaluation results based on ROUGE scores (Lin, 2004) using references collected from the user study. For each summarization test set, we have 3 machine generated summaries and 18 human summaries. We compute the ROUGE scores in a leave-1-out fashion: for each machine generated summary, we compare it against 17 out of the 18 human summaries and report the score average across the 17 runs; for each human summary, we compute the score using the other 17 as references, and report the average human summarization performance.

Evaluation results are summarized in Table 4 and Table 5, in which we report the F-measure for R-1 (unigram), R-2 (bigram) and R-SU4 (skip-bigram with maximum gap length of 4)[7], following the convention in the summarization community. Here we observe slightly different results with respect to human evaluation: for camera reviews, no significant result is observed, while HelpfulSum achieves the best R-1 score and HelpfulFilter works best regarding R-2 and R-SU4. In both cases the baseline is never the best. For movie reviews, HelpfulSum significantly outperforms the other summarizers on all ROUGE measurements, and the improvement is over 100% on R-2 and R-SU4, almost the same as human does. This is consistent with the result of pairwise comparison in that HelpfulSum works better than both HelpfulFilter and the baseline on movie reviews.

Summarizer	R-1	R-2	R-SU4
MEAD+LexRank	.333	.117	.110
HelpfulFilter	.346	**.121**	**.111**
HelpfulSum	**.350**	.110	.101
Human	.360	.138	.126

Table 4: ROUGE evaluation on camera reviews

Summarizer	R-1	R-2	R-SU4
MEAD+LexRank	.281	.044	.047
HelpfulFilter	.273	.040	.041
HelpfulSum	**.325**	**.095**	**.090**
Human	.339	.093	.093

Table 5: ROUGE evaluation on movie reviews

6 Human summary analysis

To get a comprehensive understanding of the challenges in extractive review summarization, we analyze the agreement in human summaries collected in our user study at different levels of granularity, regarding heuristics that are widely used in existing extractive summarizers.

Average word/sentence counts. Figure 4 illustrates the trend of average number of words and sentences shared by different number of users across review items for each domain. As it shows, no sentence is

[7]Because ROUGE requires all summaries to have equal length (word counts), we only consider the first 100 words in every summary.

agreed by over 10 users, which suggests that it is hard to make humans agree on the informativeness of review sentences.

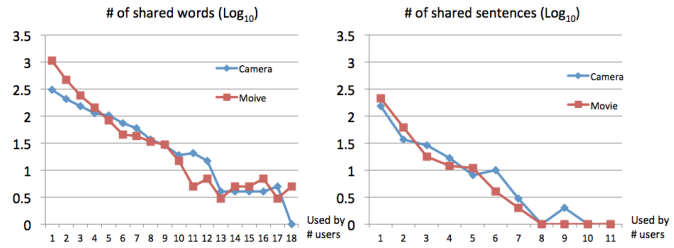

Figure 4: Average number of words (w) and sentences (s) in agreed human summaries

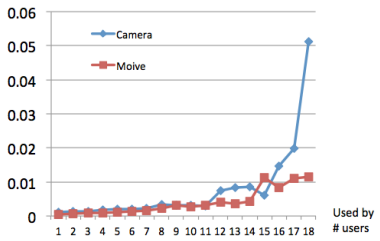

Figure 5: Average probability of words used in human summaries

Word frequency. We then compute the average probability of word (in the input) used by different number of human summarizers to see if the word frequency pattern found in news articles (words that human summarizers agreed to use in their summaries are of high frequency in the input text (Nenkova and Vanderwende, 2005)) holds for online reviews. Figure 5 confirms this. However, the average word probability is below 0.01 in those shared by 14 out of 18 summaries[8] ; the flatness of the curve seems to suggest that word frequency alone is not enough for capturing the salient information in input reviews.

KL-divergence. Another widely used heuristic in multi-document summarization is minimizing the distance of unigram distribution between the summary and the input text (Lin et al., 2006). We wonder if this applies to online review summarization. For each testing set, we group review sentences by the number of users who selected them in their summaries, and compute the KL-divergence (KLD) between each sentence group and the input. The average KL-divergence of each group across review items are visualized in Figure 6, showing that this intuition is incorrect for our review domains. Actually, the pattern is quite the opposite, especially when the number of users who share the sentences is less than 8. Thus traditional methods that aim to minimize KL-divergence might not work well for online reviews.

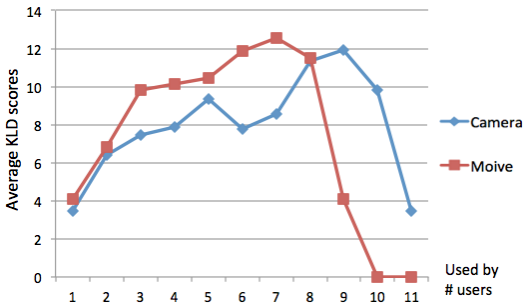

Figure 6: Average KL-Divergence between input and sentences used in human summaries

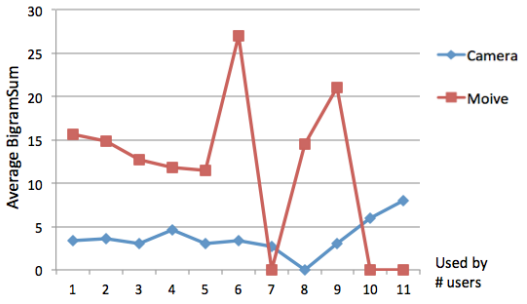

Figure 7: Average BigramSum of sentences used in human summaries

Bigram coverage. Recent studies proposed a simple but effective criteria for extractive summarization based on bigram coverage (Nenkova and Vanderwende, 2005; Gillick and Favre, 2009). The coverage of a given bigram in a summary is defined as the number of input documents the bigram appears in, and presumably good summaries should have larger sum of bigram coverage (BigramSum). However, as shown in Figure 7, this criteria might not work well in our case either. For instance, the BigramSum of the sentences that are shared by 3 human judges is smaller than those shared by 1 or 2 judges.

[8]The average probability of words used by all 4 human summarizers are 0.01 across the 30 DUC03 sets (Nenkova and Vanderwende, 2005).

7 Conclusion and future work

We propose a novel unsupervised extractive approach for summarizing online reviews by exploiting review helpfulness ratings for content selection. We demonstrate that the helpfulness metadata can not only be directly used for review-level filtering, but also be used as the supervision of sLDA for sentence scoring. This approach leverages the existing metadata of online reviews, requiring no annotation and generalizable to multiple review domains. Our experiment based on the MEAD framework shows that HelpfulFilter is preferred over the baseline (MEAD+LexRank) on camera reviews in human evaluation. HelpfulSum, which utilizes review helpfulness at both the review and sentence level, significantly outperforms the baseline in both human and automated evaluation. Our analysis on the collected human summaries reveals the limitation of traditional summarization heuristics (proposed for news articles) for being used in review domains.

In this study, we consider the ground truth of review helpfulness as the percentage of helpful votes over all votes, where the helpfulness votes could be biased in various ways (Danescu-Niculescu-Mizil et al., 2009). In the future, we would like to explore more sophisticated models of review helpfulness to eliminate such biases, or even automatic review helpfulness predictions based on just review text. We also would like to build a fully automated summarizer by replacing the review helpfulness gold-standard with automated predictions as the filtering criteria. Given the collected human summaries, we will experiment with different feature combinations for sentence scoring and we will compare our helpfulness features with other content features as well. Finally, we want to further analyze the impact of the number of human judges on our automated evaluation results based on ROUGE scores.

Acknowledgements

This research is supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A120370 to the University of Pittsburgh. The opinions expressed are those of the authors and do not necessarily represent the views of the Institute or the U.S. Department of Education. We thank Dr. Jingtao Wang and Dr. Christian Schunn for giving us suggestions on the user study design.

References

David M Blei and Jon D McAuliffe. 2010. Supervised topic models. *arXiv preprint arXiv:1003.0783*.

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

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336. ACM.

Giuseppe Carenini, Raymond T Ng, and Adam Pauls. 2006. Multi-document summarization of evaluative text. In *In Proceedings of the 11st Conference of the European Chapter of the Association for Computational Linguistics*.

Cristian Danescu-Niculescu-Mizil, Gueorgi Kossinets, Jon Kleinber g, and Lillian Lee. 2009. How opinions are received by online communities: A case study on Amazon .com helpfulness votes. In *Proceedings of the 18th International Conference on World Wide Web*, pages 141–150.

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

Anindya Ghose and Panagiotis G Ipeirotis. 2011. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering*, 23(10):1498–1512.

Dan Gillick and Benoit Favre. 2009. A scalable global model for summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing*, pages 10–18. Association for Computational Linguistics.

Aria Haghighi and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370. Association for Computational Linguistics.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Nitin Jindal and Bing Liu. 2008. Opinion spam and analysis. In *Proceedings of the international conference on Web search and web data mining*, pages 219–230.

Hyun Duk Kim and ChengXiang Zhai. 2009. Generating comparative summaries of contradictory opinions in text. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 385–394. ACM.

Soo-Min Kim, Patrick Pantel, Tim Chklovski, and Marco Pennacchiotti. 2006. Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 423–430. Association for Computational Linguistics.

Kevin Lerman and Ryan McDonald. 2009. Contrastive summarization: an experiment with consumer reviews. In *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics, companion volume: Short papers*, pages 113–116. Association for Computational Linguistics.

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.

Jure Leskovec13, Natasa Milic-Frayling, and Marko Grobelnik. 2005. Impact of linguistic analysis on the semantic graph coverage and learning of document extracts.

Jiwei Li and Sujian Li. 2012. A novel feature-based bayesian model for query focused multi-document summarization. *arXiv preprint arXiv:1212.2006*.

Chin-Yew Lin and Eduard Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th conference on Computational linguistics*, volume 1 of *COLING '00*, pages 495–501.

Chin-Yew Lin, Guihong Cao, Jianfeng Gao, and Jian-Yun Nie. 2006. An information-theoretic approach to automatic evaluation of summaries. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 463–470. Association for Computational Linguistics.

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.

Jingjing Liu, Yunbo Cao, Chin yew Lin, Yalou Huang, and Ming zhou. 2007. Low-quality product review detection in opinion summarization. In *Proceedings of the 2007 Conference on Empirical Methods in Natural Language Processing*.

Yang Liu, Xiangji Huang, Aijun An, and Xiaohui Yu. 2008. Modeling and predicting the helpfulness of online reviews. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 443–452. IEEE.

Yue Lu, Panayiotis Tsaparas, Alexandros Ntoulas, and Livia Polanyi. 2010. Exploiting social context for review quality prediction. In *Proceedings of the 19th international conference on World wide web*, pages 691–700.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Courtney Napoles, Benjamin Van Durme, and Chris Callison-Burch. 2011. Evaluating sentence compression: Pitfalls and suggested remedies. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 91–97. Association for Computational Linguistics.

Ani Nenkova and Lucy Vanderwende. 2005. The impact of frequency on summarization. *Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005-101*.

Dragomir Radev, Timothy Allison, Sasha Blair-Goldensohn, John Blitzer, Arda Celebi, Stanko Dimitrov, Elliott Drabek, Ali Hakim, Wai Lam, Danyu Liu, et al. 2004. Mead-a platform for multidocument multilingual text summarization. Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004).

Martin Riedl and Chris Biemann. 2012. How text segmentation algorithms gain from topic models. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 553–557. Association for Computational Linguistics.

Li Zhuang, Feng Jing, and Xiao-Yan Zhu. 2006. Movie review mining and summarization. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 43–50. ACM.