Helpfulness-Guided Review Summarization

Wenting Xiong
University of Pittsburgh
210 South Bouquet Street, Pittsburgh, PA 15260
wex12@cs.pitt.edu

Abstract

Review mining and summarization has been a hot topic for the past decade. A lot of effort has been devoted to aspect detection and sentiment analysis under the assumption that every review has the same utility for related tasks. However, reviews are not equally helpful as indicated by user-provided helpfulness assessment associated with the reviews. In this thesis, we propose a novel review summarization framework which summarizes review content under the supervision of automated assessment of review helpfulness. This helpfulness-guided framework can be easily adapted to traditional review summarization tasks, for a wide range of domains.

1 Introduction

Nowadays, as reviews thrive on the web, more and more people wade through these online resources to inform their own decision making. Due to the rapid growth of the review volume, the ability of automatically summarizing online reviews becomes critical to allowing people to make use of them. This makes review mining and summarization an increasingly hot topic over the past decade. Generally speaking, there are two main paradigms in review summarization. One is aspect-based opinion summarization, which aims to differentiate and summarize opinions regarding specific subject aspects. It usually involves fine-grained analysis of both review topics and review sentiment. The other is more summarization-oriented, prior work under this category either assumes a shared topic or aims to produce general summaries. In this case, the focus is the summarization, extracting salient information from reviews and organizing them properly. Compared with traditional text summarizers, sentiment-informed summarizers generally perform better as shown by human evaluation results (Carenini et al., 2006; Lerman et al., 2009).

However, one implicit assumption shared by most prior work is that all reviews are of the same utility in review summarization tasks, while reviews that comment on the same aspect and are associated with the same rating may have difference influence to users, as indicated by user-provided helpfulness assessment (e.g. "helpful" votes on Amazon.com). We believe that user-generated helpfulness votes/ratings suggest people's point of interest in review exploration. Intuitively, when users refer to online reviews for guidance, reviews that are considered helpful by more people naturally receive more attention and credit, and thus should be given more weight in review summarization. Following this intuition, we hypothesize that introducing review helpfulness information into review summarization can yield more useful review summaries.

In addition, we are also motivated by the challenges that we faced when summarizing educational peer reviews in which the review entity is also text. In the peer-review domain, traditional algorithms of identifying review aspects may suffer as reviews contain both reviewers' evaluations of a paper and reviewers' references to the paper. Such heterogeneous sources of review content bring challenges to aspect identification, and the educational perspective of peer review directly affects the characteristics of

Proceedings of the NAACL HLT 2013 Student Research Workshop, pages 77–83,
Atlanta, Georgia, 13 June 2013.

desired summaries, which has not ye been taken into consideration in any of the current summarization techniques. We expect the helpfulness assessment of peer reviews can identify important information that should be captured in peer-review summaries.

2 Related work

The proposed work is grounded in the following areas: review-helpfulness analysis, review summarization and supervised topic modeling. In this section, we will discuss existing work in the literature and explain how the proposed work relates to them.

2.1 Review-helpfulness analysis

In the literature, most researchers take a supervised approach in modeling review helpfulness. They either aggregate binary helpfulness votes for each review into a numerical score, or directly use numerical helpfulness ratings. Kim et. al (2006) took the first attempt, using regression to model review helpfulness based on various linguistic features. They reported that the combination of review length, review unigrams and product rating statistics performed best. Along this line, other studies showed the perceived review helpfulness depends not only on the review content, but also on some other factors. Ghose et. al (2008) found that the reviewer's reviewing history also matters. However, they observed that review-subjectivity, review-readability and other reviewer-related features are interchangeable for predicting review helpfulness. In addition, the empirical study on Amazon reviews conducted by Danescu-Niculescu-Mizil et. al (2009) revealed that the perceived helpfulness is also affected by how a review relates to the other reviews of the same product. However, given our goal of using review helpfulness assessment to guide summarization towards generating more useful summaries rather than to explain each individual helpfulness rating, we will ignore the interaction of helpfulness assessment among reviews of the same target.

Furthermore, the utility of features in modeling review helpfulness may vary with the review domain. Mudambi et. al (2010) showed that for product reviews, the product type moderates both the product ratings and review length on the perceived review helpfulness. For educational peer reviews, in X (2011) we showed that cognitive constructs which predict feedback implementation can further improve our helpfulness model upon general linguistic features. These findings seem to suggest that the review helpfulness model should be domain-dependent, due to the specific semantics of "helpfulness" defined in context of the domain.

2.2 Review summarization

One major paradigm of review summarization is aspect-based summarization, which is based on identifying aspects and associating opinion sentiment with them. (Although this line of work is closely related to sentiment analysis, it is not the focus of this proposed work.) While initially people use information retrieval techniques to recognize aspect terms and opinion expressions (Hu and Liu, 2004; Popescu and Etzioni, 2005), recent work seems to favor generative statistical models more (Mei et al., 2007; Lu and Zhai, 2008; Titov and McDonald, 2008b; Titov and McDonald, 2008a; Blei and McAuliffe, 2010; Brody and Elhadad, 2010; Mukherjee and Liu, 2012; Sauper and Barzilay, 2013). One typical problem with these models is that many discovered aspects are not meaningful to end-users. Some of these studies focus on distinguishing aspects in terms of sentiment variation by modeling aspects together with sentiment (Titov and McDonald, 2008a; Lu and Zhai, 2008; Mukherjee and Liu, 2012; Sauper and Barzilay, 2013). However, little attention is given to differentiating review content directly regarding their utilities in review exploration. Mukherjee and Liu (2012) attempted to address this issue by introducing user-provided aspect terms as seeds for learning review aspects, though this approach might not be easily generalized to other domains, as users' point of interest could vary with the review domain.

Another paradigm of review summarization is more summarization-oriented. In contrast, such approaches do not require the step of identifying aspects, instead, they either assume the input text share the same aspect or aim to produce general summaries. These studies are closely related to the traditional NLP task of text summarization. Generally speaking, the goal of text summarization is to retain the most important points of the input text within a shorter length. Either extractively or abstractively,

one important task is to determine the informativeness of a text element. In addition to reducing information redundancy, different heuristics were proposed within the context of opinion summarization. Stoyanov and Cardie (2008) focused on identifying opinion entities (opinion, source, target) and presenting them in a structured way (templates or diagrams). Lerman et. al (2009) reported that users preferred sentiment informed summaries based on their analysis of human evaluation of various summarization models, while Kim and Zhai (2009) further considered an effective review summary as representative contrastive opinion pairs. Different from all above, Ganesan et. al (2010) represented text input as token-based graphs based on the token order in the string. They rank summary candidates by scoring paths after removing redundant information from the graph. For any summarization framework discussed above, the helpfulness of the review elements (e.g. sentences, opinion entities, or words), which can be derived from the review overall helpfulness, captures informativeness from another dimension that has not been taken into account yet.

2.3 Supervised content modeling

As review summarization is meant to help users acquire useful information effectively, what and how to summarize may vary with user needs. To discover user preferences, Ando and Ishizaki (2012) manually analyzed travel reviews to identify the most influential review sentences objectively and subjectively, while Mukherjee and Liu (2012) extract and categorize review aspects through semi-supervised modeling using user-provided seeds (categories of terms). In contrast, we are interested in using user-provided helpfulness ratings for guidance. As these helpfulness ratings are existing meta data of reviews, we will need no additional input from users. Specifically, we propose to use supervised LDA (Blei and McAuliffe, 2010) to model review content under the supervision of review helpfulness ratings. Similar approach is widely adopted in sentiment analysis, where review aspects are learned in the presence of sentiment predictions (Blei and McAuliffe, 2010; Titov and McDonald, 2008a). Furthermore, Branavan et. al (2009) showed that joint modeling of text and user annotations benefits extractive summarization. Therefore, we hypothesize modeling review content together with review helpfulness is beneficial to review summarization as well.

3 Data

We plan to experiment on three representative review domains: product reviews, book reviews and peer reviews. The first one is mostly studied, while the later two types are more complex, as the review content consists of both reviewer's evaluations of the target and reviewer's references to the target, which is also text. This property makes review summarization more challenging.

For product reviews and book reviews, we plan to use Amazon reviews provided by Jindal and Liu (2008), which is a widely used data set in review mining and sentiment analysis. We consider the helpfulness assessment of an Amazon review as the ratio of "helpful" votes over all votes (Kim et al., 2006). For educational peer reviews, we plan to use an annotated corpus (Nelson and Schunn, 2009) collected from an online peer-review reciprocal system, which we used in our prior work (Xiong and Litman, 2011). Two experts (a writing instructor and a content instructor) were asked to rate the helpfulness of each peer review on a scale from one to five (Pearson correlation r = 0.425, $p \leq 0.01$). For our study, we consider the average ratings given by the two experts (which roughly follow a normal distribution) as the gold standard of review helpfulness ratings. To be consistent with the other review domains, we normalize peer-review helpfulness ratings in the range between 0 and 1.

4 Proposed work

The proposed thesis work consists of three parts: 1) review content analysis using user-provided helpfulness ratings, 2) automatically predicting review helpfulness and 3) a helpfulness-guided review summarization framework.

4.1 Review content analysis

Before advocating the proposed idea, we would test our two hypothesis: 1) user-provided review helpfulness assessment reflects review content difference. 2) Considering review content in terms of **internal content** (e.g. reviewers' opinions) vs. **external content** (e.g. book content), the internal content

influences the perceived review helpfulness more than the external content.

We propose to use two kind of instruments, one is Linguistic Inquiry Word Count (LIWC)[1], which is a manually created dictionary of words; the other is the set of review topics learned by Latent Dirichlet Allocation (LDA) (Blei et al., 2003; Blei and McAuliffe, 2010). LIWC analyzes text input based on language usages both syntactically and semantically, which reveals review content patterns at a high level; LDA can be used to model sentence-level review topics which are domain specific.

For the LIWC-based analysis, we test whether each category count has a significant effect on the numerical helpfulness ratings using paired T-test. For LDA-based analysis, we demonstrate the difference by show how the learned topics vary when helpfulness information is introduced as supervision. Specifically, by comparing the topics learned from the unsupervised LDA and those learned from the supervised LDA (with helpfulness ratings), we expect to show that the supervision of helpfulness ratings can yield more meaningful aspect clusters.

It is important to note that in both approaches a review is considered as a bag of words, which might be problematic if the review has both internal and external content. Considering this, we hypothesize that the content difference captured by user-provided helpfulness ratings is mainly in the reviewers' evaluation rather than in the content of external sources (hypothesis 2). We plan to test this hypothesis on both book reviews and peer reviews by analyzing review content in two conditions: in the first condition (the control condition), all content is preserved; in the second condition, the external content is excluded. If we observe more content variance in the second condition than the first one, the second hypothesis is true. Thus we will separate review internal and external content in the later summarization step. For simplification, in the second condition, we only consider the topic words of the external content; we plan to use a corpus-based approach to identify these topic terms and filter them out to reduce the impact of external content.

[1]Url: http://www.liwc.net. We are using LIWC2007.

4.2 Automated review helpfulness assessment

Considering how review usefulness would be integrated in the proposed summarization framework, we propose two models for predicting review helpfulness at different levels of granularity.

A discriminative model to learn review global helpfulness. Previously we (2011) built a discriminative model for predicting the helpfulness of educational peer reviews based on prior work of automatically predicting review helpfulness of product reviews (Kim et al., 2006). We considered both domain-general features and domain-specific features. The domain-general features include structure features (e.g. review length), semantic features, and descriptive statistics of the product ratings (Kim et al., 2006); the domain-specific features include the percentage of external content in reviews and cognitive and social science features that are specific to the peer-review domain. To extend this idea to other types of reviews: for product reviews, we consider product aspect-related terms as the topic words of the external content; for book reviews, we take into account author's profile information (number of books, the mean average book ratings). As we showed that replacing review unigrams with manually crafted keyword categories can further improve the helpfulness model of peer reviews, we plan to investigate whether review unigrams are generally replaceable by review LIWC features for modeling review helpfulness.

A generative model to learn review local helpfulness. In order to utilize user-provided helpfulness information in a decomposable fashion, we propose to use sLDA (Blei and McAuliffe, 2010) to model review content with review helpfulness information at the review level, so that the learned latent topics will be predictive of review helpfulness. In addition to evaluating the model's predictive power and the quality of the learned topics, we will also investigate the extent to which the model's performance is affected by the size of the training set, as we may need to use automatically predicted review helpfulness instead, if user-provided helpfulness information is not available.

4.3 Helpfulness-guided review summarization

In the proposed work, we plan to investigate various methods of supervising an extractive review summarizer using the proposed helpfulness models. The simplest method (M1) is to control review helpfulness of the summarization input by removing reviews that are predicted of low helpfulness. A similar method (M2) is to use post-processing rather than pre-processing – reorder the selected summary candidates (e.g. sentences) based on their predicted helpfulness. The helpfulness of a summary sentence can be either inferred from the local-helpfulness model (sLDA), or aggregated from review-level helpfulness ratings of the review(s) from which the sentence is extracted. The third one (M3) works together with a specific summarization algorithm, interpolating traditional informativeness assessment with novel helpfulness metrics based on the proposed helpfulness models.

For demonstration, we plan to prototype the proposed framework based on MEAD* (Carenini et al., 2006), which is an extension of MEAD (an open-source framework for multi-document summarization (Radev et al., 2004)) for summarizing evaluative text. MEAD* defines sentence informativeness based on features extracted through standard aspect-based review mining (Hu and Liu, 2004). As a human-centric design, we plan to evaluate the proposed framework in a user study in terms of pairwise comparison of the reviews generated by different summarizers (M1, M2, M3 and MEAD*). Although fully automated summarization metrics are available (e.g. Jensen-Shannon Divergence (Louis and Nenkova, 2009)), they favor summaries that have a similar word distribution to the input and thus do not suit our task of review summarization.

To show the generality of the proposed ideas, we plan to evaluate the utility of introducing review helpfulness in aspect ranking as well, which is an important sub-task of review opinion analysis. If our hypothesis (1) is true, we would expect aspect ranking based on helpfulness-involved metrics outperforming the baseline which does not use review helpfulness (Yu et al., 2011). This evaluation will be done on product reviews and peer reviews, as the previous work was based on product reviews, while peer reviews tend to have an objective aspect ranking (provided by domain experts).

5 Contributions

The proposed thesis mainly contributes to review mining and summarization.

1. Investigate the impact of the source of review content on review helpfulness. While a lot of studies focus on product reviews, we based our analysis on a wider range of domains, including peer reviews, which have not been well studied before.

2. Propose two models to automatically assess review helpfulness at different levels of granularity. While the review-level global helpfulness model takes into account domain-specific semantics of helpfulness of reviews, the local helpfulness model learns review helpfulness jointly with review topics. This local helpfulness model allows us to decompose overall review helpfulness into small elements, so that review helpfulness can be easily combined with metrics of other dimensions in assessing the importance of summarization candidates.

3. Propose a user-centric review summarization framework that utilizes user-provided helpfulness assessment as supervision. Compared with previous work, we take a data driven approach in modeling review helpfulness as well as helpfulness-related topics, which requires no extra human input of user-preference and can be adapted to typical review summarization tasks such as aspect selection/ranking, summary sentence ordering, etc.

References

M. Ando and S. Ishizaki. 2012. Analysis of travel review data from readers point of view. *WASSA 2012*, page 47.

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

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

SRK Branavan, H. Chen, J. Eisenstein, and R. Barzilay. 2009. Learning document-level semantic properties

from free-text annotations. *Journal of Artificial Intelligence Research*, 34(2):569.

S. Brody and N. Elhadad. 2010. An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 804–812. Association for Computational Linguistics.

G. Carenini, R. Ng, and A. 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*. Citeseer.

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 WWW*, pages 141–150.

Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. 2010. Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 340–348, Stroudsburg, PA, USA. Association for Computational Linguistics.

Anindya Ghose and Panagiotis G. Ipeirotis. 2008. Estimating the socio-economic impact of product reviews. In *NYU Stern Research Working Paper CeDER*.

M. Hu and B. 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.

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

H.D. Kim and C.X. 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 (EMNLP2006)*, pages 423–430, Sydney, Australia, July.

K. Lerman, S. Blair-Goldensohn, and R. 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. Association for Computational Linguistics.

Annie Louis and Ani Nenkova. 2009. Automatically evaluating content selection in summarization without human models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 306–314. Association for Computational Linguistics.

Y. Lu and C. Zhai. 2008. Opinion integration through semi-supervised topic modeling. In *Proceedings of the 17th international conference on World Wide Web*, pages 121–130. ACM.

Q. Mei, X. Ling, M. Wondra, H. Su, and C.X. Zhai. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th international conference on World Wide Web*, pages 171–180. ACM.

S.M. Mudambi and D. Schuff. 2010. What makes a helpful online review? a study of customer reviews on amazon. com. *MIS quarterly*, 34(1):185–200.

A. Mukherjee and B. Liu. 2012. aspect extraction through semi-supervised modeling. In *Proceedings of 50th anunal meeting of association for computational Linguistics (acL-2012)(accepted for publication)*.

Melissa M. Nelson and Christian D. Schunn. 2009. The nature of feedback: how different types of peer feedback affect writing performance. In *Instructional Science*, volume 37, pages 375–401.

A.M. Popescu and O. Etzioni. 2005. Extracting product features and opinions from reviews. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 339–346. Association for Computational Linguistics.

D. Radev, T. Allison, S. Blair-Goldensohn, J. Blitzer, A. Celebi, S. Dimitrov, E. Drabek, A. Hakim, W. Lam, D. Liu, et al. 2004. Mead-a platform for multidocument multilingual text summarization. In *Proceedings of LREC*, volume 2004.

C. Sauper and R. Barzilay. 2013. Automatic aggregation by joint modeling of aspects and values. *Journal of Artificial Intelligence Research*, 46:89–127.

V. Stoyanov and C. Cardie. 2008. Topic identification for fine-grained opinion analysis. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 817–824. Association for Computational Linguistics.

I. Titov and R. McDonald. 2008a. A joint model of text and aspect ratings for sentiment summarization. *Urbana*, 51:61801.

I. Titov and R. McDonald. 2008b. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th international conference on World Wide Web*, pages 111–120. ACM.

Wenting Xiong and Diane Litman. 2011. Automatically predicting peer-review helpfulness. In *Proceedings of*

the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 502–507.

J. Yu, Z.J. Zha, M. Wang, and T.S. Chua. 2011. Aspect ranking: identifying important product aspects from online consumer reviews. *Computational Linguistics*, pages 1496–1505.