Sentence Retrieval with Sentiment-specific Topical Anchoring for Review Summarization

Jiaxing Tan
City University of New York, New York, USA
jtan@gradcenter.cuny.edu

Alexander Kotov
Wayne State University, Detroit, USA
kotov@wayne.edu

Rojiar Pir Mohammadiani
K. N. Toosi University of Technology, Tehran, Iran
rpirmohamadiani@mail.kntu.ac.irg

Yumei Huo
City University of New York, New York, USA
yumei.huo@csi.cuny.edu

ABSTRACT

We propose Topic Anchoring-based Review Summarization (TARS), a two-step extractive summarization method, which creates review summaries from the sentences that represent the most important aspects of a review. In the first step, the proposed method utilizes Topic Aspect Sentiment Model (TASM), a novel sentiment-topic model, to identify aspects of sentiment-specific topics in a collection of reviews. The output of TASM is utilized in the second step of TARS to rank review sentences based on how representative of the most important review aspects their words are. Qualitative and quantitative evaluation of review summaries using two collections indicate the effectiveness of structuring review summaries around aspects of sentiment-specific topics.

CCS CONCEPTS

•Information systems →Summarization;

KEYWORDS

Review Summarization, Opinion Mining, Topic Models

ACM Reference format:

Jiaxing Tan, Alexander Kotov, Rojiar Pir Mohammadiani, and Yumei Huo. 2017. Sentence Retrieval with Sentiment-specific Topical Anchoring for Review Summarization. In *Proceedings of CIKM'17, November 6–10, 2017, Singapore., ,* 4 pages.

DOI: https://doi.org/10.1145/3132847.3133153

1 INTRODUCTION

The past decade has witnessed the emergence and tremendous increase in popularity of on-line consumer review platforms for a wide variety of products and services. However, large volume of reviews published on these platforms can make it difficult for users to quickly form a "big picture" of the overall sentiment towards different aspects of a product or service based on its reviews. Review summarization methods address this issue by distilling the content of reviews to generate their more concise versions.

Extractive summarization approaches generate a summary by selecting fragments (e.g. sentences) from the original review. In particular, graph-based summarization approaches [1, 3, 11] first construct a graph, in which the nodes are review sentences and the weighted edges represent the degree of similarity or content overlap between the sentences. Once the graph is constructed, eigenvector centrality measures, such as PageRank, can be used to rank the sentences. Extractive review summarization methods can also leverage information retrieval techniques [7, 10]. Specifically, the method proposed in [7] generates a summary by ranking review sentences based on how explanatory they are, which is captured by three features: sentence length, popularity and discriminativeness of its words. Review summaries can also be created by retrieving questions from on-line question answering platforms [10].

In this paper, we propose Topic Anchoring-based Review Summarization (TARS), a two-step extractive summarization method, which is based on the intuition that a comprehensive review summary should:

- represent all major topics (e.g. CPU, battery) and specific aspects of these topics (e.g. CPU performance, battery capacity) discussed in a review of a particular product (e.g. laptop);
- reflect the polarity of opinions towards these aspects and topics as well as the main reasons for these opinions.

Furthermore, while many different aspects of a product or service can be discussed in its reviews, review summaries should prioritize the most important and popular aspects, since they provide a strong and reliable feedback to both manufacturers and consumers. Important aspects are the aspects that constitute the majority of review content, whereas popular aspects are the aspects that are discussed in a large number of reviews.

To generate review summaries that satisfy the above desiderata, TARS utilizes the output of Topic Aspect Sentiment Model (TASM), a novel topic model for opinion mining. Specifically, TASM is used in the first step of TARS to obtain the word distributions corresponding to sentiment-specific aspects of the major topics discussed in a collection of reviews. Topic models, such as Latent Dirichlet Allocation [2], are probabilistic generative models, which can automatically detect the latent structure in a document collection in the form of clusters of semantically related words (topics) that are shared

CIKM'17, November 6–10, 2017, Singapore.
 ISBN 978-1-4503-4918-5/17/11…$15.00
DOI: https://doi.org/10.1145/3132847.3133153

between documents. In opinion analysis, topic models are used to extract an unstructured set of fine-grained sentiment-specific topics [6, 9], a set of sentiment-specific topics for each demographic group of review authors [13] or a hierarchy of sentiment-specific topics [8] from a collection of reviews. TASM is different from other topic models for opinion mining in that it considers each review as a mixture of aspects of sentiment-specific topics rather than sentiment-specific topics, which allows TASM to better capture a unique combination of topics, aspects and opinions in each individual review. The key idea behind TARS is that *summaries should be structured around aspects of sentiment-specific topics, which serve as "anchors" for selecting review sentences.* Therefore, word distributions associated with aspects of sentiment-specific topics are used in the second step of TARS to rank review sentences according to popularity and representativeness of review aspects they are covering, which results in summaries that capture all important points in reviews.

The main contributions of this work are two-fold:

(1) a novel topic model for opinion mining that identifies a set of fine-grained topics corresponding to aspects of sentiment-specific topics in a given collection of reviews;
(2) the first ranking based extractive summarization method, which leverages the output of sentiment-topic model to create the summary of a review by selecting the sentences based on how well they cover the most important aspects of a review. Previously proposed extractive summarization methods are designed for generic documents and do not take into account sentiment-specific topics or aspects.

2 METHOD

In this section, we discuss the details of the two stages in the proposed method.

2.1 Topic Aspect Sentiment Model

TASM considers each review as a 3-level mixture. The first level of the mixture corresponds to a distribution over the major topical themes in a collection. The second level of the mixture corresponds to a distribution over aspects for each major topical theme. Finally, the third level corresponds to a distribution over sentiments for each aspect. TASM models reviews according to the following probabilistic generative process, which is illustrated in Figure 1:

(1) draw a multinomial distribution over vocabulary $\phi_b \sim Dir(\beta)$ for the background topic
(2) draw a multinomial distribution over vocabulary $\phi_{zys} \sim Dir(\beta_s)$ for each combination of K topics, Y aspects and S sentiments
(3) for each review d of M reviews in the collection:
    (a) draw $\pi_d \sim Beta(\omega)$, a binomial distribution determining the mixture of background and non-background topics
    (b) draw $\theta_d \sim Dir(\alpha)$, a multinomial distribution over non-background topics
    (c) for each non-background topic z, draw $\Omega_{zd} \sim Dir(\tau)$, a multinomial distribution over aspects

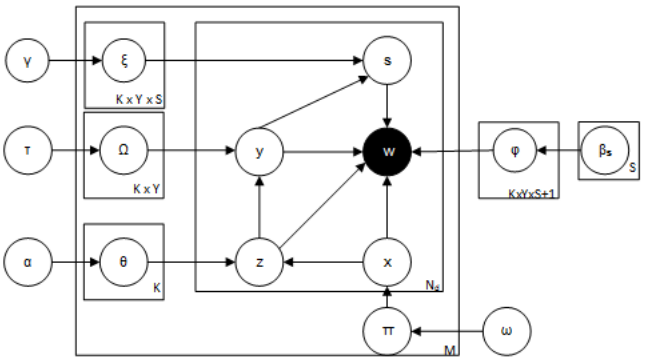

Figure 1: Generative process of TASM in plate notation

(d) for each non-background topic z and aspect y, draw $\xi_{yzd} \sim Dir(\gamma)$, a multinomial distribution over sentiments
(e) for each word position w_i of N_d word positions in d:
    (i) sample a Bernoulli variable x from π_d
    (ii) if $x = 0$, sample w_i from ϕ_b
    (iii) if $x = 1$:
        (A) sample a topic $z \sim \theta_d$
        (B) sample an aspect $y \sim \Omega_{zd}$
        (C) sample a sentiment $s \sim \xi_{zyd}$
        (D) sample a word $w_i \sim \phi_{zys}$

Very frequent words (e.g. stopwords) are not representative of review aspects and, as such, should have little impact on selecting review sentences for a summary. The background topic absorbs such words, thus helping to improve the quality of summaries.

Posterior inference of TASM parameters is performed via Gibbs sampling. In each state of the Markov chain of Gibbs sampler, each word i in review d is assigned to the background topic with probability:

$$p(x_{d,i} = 0|\vec{w}, \vec{x}) \propto \frac{n^{\neg d,i}_{d,.,.,.,.,.,x=0} + \omega_{x=0}}{n^{\neg d,i}_{d,.,.,.,.,.,.} + \omega_{x=0} + \omega_{x=1}} \cdot \frac{n^{\neg d,i}_{.,.,.,.,.,w,x=0} + \beta}{n^{\neg d,i}_{.,.,.,.,.,.,x=0} + V\beta} \quad (1)$$

or to the non-background topic k, aspect y and sentiment s with probability:

$$p(x_{d,i} = 1, z_{d,i} = k, y_{d,i} = y, s_{d,i} = s|\vec{w}, \vec{x}, \vec{z}, \vec{y}, \vec{s}) \propto \frac{n^{\neg d,i}_{d,k,.,.,.,.,x=1} + \omega_{x=1}}{n^{\neg d,i}_{d,.,.,.,.,.,.} + \omega_{x=0} + \omega_{x=1}} \cdot \frac{n^{\neg d,i}_{.,k,y,s,w,x=1} + \beta^w_s}{n^{\neg d,i}_{.,k,y,s,.,x=1} + \sum_{j=1}^{V} \beta^{w_j}_s} \cdot \frac{n^{\neg d,i}_{d,k,y,.,.,.,x=1} + \tau}{n^{\neg d,i}_{d,k,.,.,.,.,x=1} + Y\tau} \cdot \frac{n^{\neg d,i}_{d,k,y,s,.,.,x=1} + \gamma}{n^{\neg d,i}_{d,k,y,.,.,.,x=1} + 3\gamma} \cdot (n^{\neg d,i}_{d,k,.,.,.,.,x=1} + \alpha) \quad (2)$$

where $n^{\neg d,i}_{d,k,y,s,.,x=1}$ is the number of times non-background topic k, aspect y and sentiment s have been assigned to words in review d, $n^{\neg d,i}_{.,k,y,s,w,x=1}$ is the number of times topic k, aspect y and sentiment s have been assigned to word w in all reviews (both excluding the assignment to the ith word in d) and $n^{\neg d,i}_{d,.,.,.,.,.,.} = |d| - 1$.

To set non-uniform Dirichlet priors β_s^w for aspects of sentiment-specific topics, we compiled a sentiment lexicon, which combines the PARADIGMhasm sentiment lexicon [8] consisting of 31 positive and 33 negative words, MPQA [12] sentiment lexicon consisting of 2718 positive words and 4911 negative words and the sentiment lexicon from [5] consisting of 2006 positive words and 4783 negative words. The compiled lexicon is used to set the priors as follows:

$$\beta_s^w = \begin{cases} \beta_s, & \text{if } w \in dict(s) \\ \beta_{-s}, & \text{if } w \in dict(-s) \\ \beta & \text{otherwise} \end{cases} \quad (3)$$

where $0 < \beta_{-s} < \beta < \beta_s < 1$, $dict(s)$ is a dictionary for sentiment s in the compiled lexicon and $-s$ is the opposite sentiment.

2.2 Sentence Ranking Method

The word distributions for aspects of sentiment-specific topics discovered by TASM are used in the second step of TARS to rank each sentence in a given review based on:

(1) how *distinct* its individual words are to a particular review aspect;
(2) how *representative* it is as a whole of a particular review aspect;
(3) how *important* is the aspect that the sentence corresponds to in a review.

These criteria are applied at different levels: the first criteria quantifies the property of each individual word in a sentence, the second criteria quantifies the property of an entire sentence, while the third criteria quantifies the property of a review aspect that a sentence corresponds to. According to these criteria, the review sentences that are ranked high (and thus are likely to be selected for a summary) are *highly representative of important review aspects*. ψ_{zysd}, the popularity of aspect y of topic z specific to sentiment s in a given review d can be quantified as the proportion of review words that are sampled from ϕ_{zys}, which can be calculated based on the results of posterior inference of TASM parameters as follows:

$$\psi_{zysd} = p(z, y, s|d) = p(z|d)p(y|z, d)p(s|y, z, d) = \theta_d \Omega_{zd} \xi_{zyd} \quad (4)$$

The representativeness score of sentence t with respect to aspect y of topic z with sentiment s is determined by aggregating the aspect distinctness scores of its individual words:

$$RS(t; z, y, s) = \frac{\sum_{w \in t} DS(w; z, y, s)}{|t|} \quad (5)$$

where $|t|$ is the sentence length (total number of words in the sentence), which acts as a normalizer. In the context of the proposed ranking method, distinctness of a word not only implies that a word is important in a particular aspect of sentiment-specific topic (i.e. it is assigned a high probability in the distribution corresponding to this aspect), but also that a word is important only in that particular aspect. Specifically, word distributions for aspects of sentiment-specific topics obtained by TASM are utilized to calculate the distinctness score of word w with respect to aspect y of topic z with sentiment s as follows:

$$DS(w; z, y, s) = \frac{p(w|z, y, s)}{1 + p(w|\bar{z}, \bar{y}, \bar{s})} \quad (6)$$

where $p(w|\bar{z}, \bar{y}, \bar{s})$ represents cumulative importance of word w in all aspects of sentiment-specific topics other than z, y and s and is calculated as:

$$p(w|\bar{z}, \bar{y}, \bar{s}) = \sum_{\forall(z_i, y_i, s_i): z_i \neq z, y_i \neq y, s_i \neq s} p(w|z_i, y_i, s_i) \quad (7)$$

The word distinctness score is based on a simple intuition that non-representative words are important in many sentiment-specific topical aspects, whereas representative words are important in only a few or one aspect.

The proposed ranking based extractive summarization method is summarized in Algorithm 1.

Algorithm 1 Sentence Ranking Algorithm

Input: RS, a set of sentences in review d
Input: L, number of sentences in a summary
Input: ψ_{zysd}, importance of aspects in review d
Output: SS, a vector of sentences in a summary of d

1: $SA \leftarrow queue(sort(\psi_{zysd}))$
2: $i \leftarrow 0$
3: $SS \leftarrow \emptyset$
4: **while** $i \leq L$ **do**
5: $(z, y, s) \leftarrow SA.pop\_first()$
6: $t_{zys} = \underset{t \in RS}{\text{argmax}}\, RS(t; z, y, s)$
7: $SS.append(t_{zys})$
8: $RS \leftarrow RS \setminus t_{zys}$
9: $SA.push\_back((z, y, s))$
10: $i \leftarrow i + 1$
11: **end while**

First, the aspects of sentiment-specific topics in ψ_{zysd} are sorted in descending order of their probabilities and added to the queue SA (line 1). A summary is constructed by repeatedly selecting the aspects from SA and finding the most representative sentence for each aspect among the remaining sentences in RS, until L sentences have been selected (lines 4-10).

3 EXPERIMENTS

3.1 Datasets and experimental design

We evaluated our proposed method using two datasets. The first dataset is a hotel review dataset used in [13]. The second dataset is the one used in [4]. Specifically, two experiments were conducted. In the first experiment, we asked human judges to compare the quality of the summaries created by our method and Lexrank[3], a popular extractive summarization method. First, we randomly picked 100 pieces of reviews, each having more than 800 words, from the hotel review dataset and generated their summaries using the two methods. Then we asked human judges to decide which summary is better by providing a score from 1 to 5. The score of 1 indicates that TARS summary is much better, the score of 5 indicates that Lexrank summary is much better, while the score of

	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-SU4
TARS	**0.3146**	**0.0693**	**0.3104**	**0.1252**
TARS-NN	0.3122	0.0544	0.2836	0.1146
Lexrank	0.3040	0.0524	0.2927	0.1233
R-SumWordLR	0.2471	0.0392	0.2272	0.0823
TRS-LDA	0.3078	0.0501	0.2907	0.1121

Table 1: ROUGE recall for TARS, its variants and Lexrank

3 indicates there is no difference between the summaries generated by the methods.

In the second experiment, we used ROUGE, an automatic evaluation method, which measures the overlap between the generated summaries and the gold standard summaries. In this experiment, we compare our method with several baselines:

- **TARS-NN:** TARS without normalization (no denominator in Eq. 5);
- **Lexrank**: a popular extractive summarization method [3];
- **R-SumWordLR**: revised SumWordLR from [7] using the topics generated by TASM. $P(w|B=0)$, posterior probability that word w is not explanatory, directly comes from the background topic, while $P(w|B=1)$, posterior probability that word w is explanatory, is calculated as:

$$P(w|B=1) = \sum_{z \in K, y \in Y, s \in S} p(w|z,y,s)p(z,y,s) \quad (8)$$

- **TRS-LDA:** topic-based review summarization (TRS) using LDA. Reviews are generated by repeatedly selecting the most representative sentences for review topics according to their importance.

The results for TARS were obtained by setting $K = 20$, $Y = 3$, $\gamma = 0.1$, $\tau = 0.1$, $\omega = 0.1$, $\alpha = 50/K$, $\beta = 0.01$, $\beta_s = 0.15$ and $\beta_{-s} = 0.005$ for TASM. The results for TARS-LDA were obtained by setting $K = 20$, $\alpha = 50/K$ and $\beta = 0.01$ for LDA.

3.2 Results

For the first experiment, the average score over all reviews is 2.24, which indicates that human judges favored the reviews created by TARS over those created by Lexrank in majority of cases. The results of the second experiment are summarized in Table 1. We measured the quality of generated summaries in terms of ROUGE metrics: based on unigrams (ROUGE-1), bigrams (ROUGE-2), longest common subsequence (ROUGE-L) and unigram and skip-bigrams separated by up to four words (ROUGE-SU4). Four major conclusions can be drawn from the results in Table 1. First, TARS outperforms all its variants and Lexrank in terms of recall for all four ROUGE metrics. Second, ranking sentences according to how characteristic they are of popular review aspects results in better summaries than ranking them by their explanatoriness as in [7]. Third, reviews that are structured around sentiment-specific topics are better than review that are structured around LDA topics. Finally, normalizing the sentence representativeness score allows to slightly improve the quality of summaries.

As a final experiment, we measured sensitivity of TARS performance to the number of topics in TASM, which is illustrated in Figure 2. As follows from Figure 2, the quality of summaries generated by TARS depends on the number of topics for TASM. Therefore, effective practical application of the proposed method requires preliminary analysis of the corpus. This issue can be addressed by developing a non-parametric version of TASM, which we leave as future work.

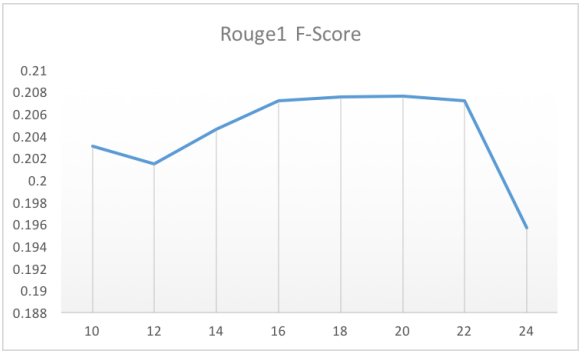

Figure 2: ROUGE-1 F-score by varying the number of topics

4 CONCLUSION

In this paper, we proposed an extractive summarization method based on ranking review sentences according to how characteristic they are of important review aspects identified by a sentiment-topic model. Experimental evaluation involving human judges and automatic metrics indicates that structuring reviews around aspects of sentiment-specific topics is a more effective strategy than other ranking heuristics.

REFERENCES

[1] Mohammed Al-Dhelaan. 2015. StarSum: A Simple Star Graph for Multi-document Summarization. In *Proceedings of ACM SIGIR*. 715–718.

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

[3] Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research* 22 (2004), 457–479.

[4] Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. 2010. Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In *Proceedings of COLING*. 340–348.

[5] Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of ACM SIGKDD*. 168–177.

[6] Yohan Jo and Alice H Oh. 2011. Aspect and sentiment unification model for online review analysis. In *Proceedings of ACM WSDM*. 815–824.

[7] Hyun Duk Kim, Malu G Castellanos, Meichun Hsu, ChengXiang Zhai, Umeshwar Dayal, and Riddhiman Ghosh. 2013. Ranking explanatory sentences for opinion summarization. In *Proceedings of ACM SIGIR*. 1069–1072.

[8] Suin Kim, Jianwen Zhang, Zheng Chen, Alice Oh, and Shixia Liu. 2013. A hierarchical aspect-sentiment model for online reviews. In *Proceedings of AAAI*. 526–533.

[9] Chenghua Lin and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of ACM CIKM*. 375–384.

[10] Mengwen Liu, Yi Fang, Dae Hoon Park, Xiaohua Hu, and Zhengtao Yu. 2016. Retrieving Non-Redundant Questions to Summarize a Product Review. In *Proceedings of ACM SIGIR*. 385–394.

[11] Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into texts. In *Proceedings of EMNLP*. 404–411.

[12] Peter D Turney and Michael L Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems* 21, 4 (2003), 315–346.

[13] Zaihan Yang, Alexander Kotov, Aravind Mohan, and Shiyong Lu. 2015. Parametric and non-parametric user-aware sentiment topic models. In *Proceedings of ACM SIGIR*. 413–422.