Ranking Sentiment Explanations for Review Summarization Using Dual Decomposition

Lei Fang, Qiao Qian, Minlie Huang and Xiaoyan Zhu
State Key Laboratory of Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology
Tsinghua University, Beijing 100084, PR China
fang-l10@mails.tsinghua.edu.cn, qianqiaodecember29@126.com,
aihuang@tsinghua.edu.cn, zxy-dcs@tsinghua.edu.cn

ABSTRACT

For online reviews, sentiment explanations refer to the sentences that may suggest detailed reasons of sentiment, which are very important for applications in review mining like opinion summarization. In this paper, we address the problem of ranking sentiment explanations by formulating the process as two subproblems: sentence informativeness ranking and structural sentiment analysis. Tractable inference in joint prediction is performed through dual decomposition. Preliminary experiments on publicly available data demonstrate that our approach obtains promising performance.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information filtering*; I.2.7 [**Artificial Intelligence**]: Natural language processing—*Text Analysis*

General Terms

Algorithms, Experimentation

Keywords

Opinion Mining; Sentiment Explanation; Dual Decomposition

1. INTRODUCTION

With the ongoing increasing amount of user-generated reviews on the web, many people consider online reviews as guidelines for decision making. However, few websites provide brief summaries, which makes it difficult for users to find what they focus on, particularly when the size of reviews is very large. On the other hand, for a single review, not every part is equally informative. It would be important to highlight the informative part of each review before review summarization. We term the informative part as "sentiment explanations".

From our point of view, "sentiment explanations" may be several sentences that suggest the detailed reasons of sentiment. Sentiment explanations are valuable for abstracting a single review, which would also benefit the performance of summarizing a collection of reviews. We propose that good sentiment explanations of a single review should have the following properties:

1) from the summarization perspective, they best represent the content of original review.
2) from the sentiment perspective, they best represent the key opinion of original review.

The second property ensures that sentiment explanations should represent original reviews in terms of sentiment polarity, because a review might consist of various opinions. It is an advanced property compared with traditional single or multi document summarization.

In this paper, we propose to rank the sentences of a single review such that sentiment explanations rank higher. We formulate the ranking process as two subproblems: sentence informativeness ranking and multi-level sentiment analysis, which also echo the two properties for sentiment explanations. For sentence informativeness ranking, we train a simple ranking model from unlabeled data with several heuristic rules; for multi-level sentiment analysis, we employ the approach proposed by Yessenalina *et al.*[16] which aims to select sentences that best represent the original review in terms of polarity. Tractable inference in joint models is performed through dual decomposition [14]. Preliminary experiments on publicly available data set demonstrate that our approach of joint modeling obtains promising performance.

2. RELATED WORK

Recent years, there has been many studies focused on sentiment analysis [12]. Pang and Lee [11] and Yessenalina *et al.*[16] shown that not every part of the review was equally informative, they obtained improved sentiment classification performance as they considered that subjective part was more important for inferring the review rating.

For review summarization, generally it is considered as a sentence or review selection problem [1, 7]. Other studies performed review summarization in cascade approaches[3, 4, 10, 18] with first opinion extraction and then document summarization. However, these approaches don't highlight on sentiment explanations.

One similar work with this paper is [6], they scored the explanatoriness for each sentence, and then ranked explanatory sentences for opinion summarization. Though their approach was unsupervised, their formulation was based on the assumption that exiting technique can be used to classify the aspect and sentiment of each review. Our setting is more fundamental and our approach is closer to pragmatic needs.

CIKM'14, November 3–7, 2014, Shanghai, China.

ACM 978-1-4503-2598-1/14/11 ...$15.00.
http://dx.doi.org/10.1145/2661829.2662048.

Symbol	Description
x	a review document
$\|x\|$	number of sentences in x
y	sentiment polarity
x^j	a review sentence
A	a set of aspect seeds
a	a aspect label
V	vocabulary

Table 1: Basic Notations

3. JOINT SENTENCE RANKING AND SENTIMENT ANALYSIS

We first propose two subproblems for each property of sentiment explanation: sentence informativeness ranking and multi-level sentiment analysis. After that, joint inference of two subproblems will benefit the sentence informativeness ranking model such that sentiment explanations rank higher. For our task here, tractable inference in joint prediction is performed through dual decomposition [14].

Dual decomposition is a general approach for combinatorial optimization, with each sub-problem can be solved separately. With the help of dual decomposition, it makes the task of sentiment explanation ranking much more easier. We first present the setting for sentence informativeness ranking and multi-level sentiment analysis, respectively; then we give details for joint inference on a new review using dual decomposition. Table 1 presents notations we will use throughout this paper.

3.1 Sentence Informativeness Ranking

We propose the following heuristic rules for sentence informativeness ranking:

- the sentence would rank higher if it contains more opinion words[1];
- the sentence would rank higher if it contains more aspect words.

Aspects can be considered as certain properties of a product or service. For example, the aspects are "story", "music", "acting", "picture" and "director" for movie reviews; "taste", "ambience", "service", "price" and "location" for restaurant reviews. We extract aspect terms using a bootstrapping algorithm based on Chi-Square (χ^2) statistics shown in Algorithm 1, which is similar with[15].

The χ^2 statistic to compute the dependencies between word v and aspect a_j is

$$\chi^2(v, a_j) = \frac{C \times (C_1C_4 - C_2C_3)}{(C_1{+}C_3) \times (C_2{+}C_4) \times (C_1{+}C_2) \times (C_3{+}C4)};$$

where C_1 is the number of times v occurs in sentences with aspect label a_j, C_2 is the number of times v occurs in sentences not labeled with a_j, C_3 is the number of sentences with aspect a_j but do not contain v, C_4 is the number of sentences that neither belong to aspect a_j nor contain word v, and C is the total number of word occurrences.

After extraction of aspect words, we generate the rank of each sentence for a collection of unlabeled sentences based on the aforementioned two rules. Then, we are able to train a ranking model using some learning to rank [8] techniques. In this work, we choose a pairwise ranking approach: SVM^{rank}[5], and use bag-of-words features to train the ranking model.

[1] We use the sentiment lexicon from `http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html`

Algorithm 1 Bootstrapping Framework.

Input:
A collection of review sentences, $X = \{x^1, x^2, \ldots\}$;
A collection of aspect seeds sets $A_1, A_2, \ldots$;
Selection threshold n, iteration step limit l;
Output:
Extended aspect word sets $T_1, T_2, \ldots$;

Initialize $T_i = A_i$ for all aspects
repeat
  for all sentence $x^i \in X$ **do**
    Match aspect words for x^i, and record the matching hits for aspect a_j in $Count(j)$
    Assign aspect label a_j to x^i if $a_j = \underset{j}{\mathrm{argmax}} Count(j)$
  end for
  for all aspect a_j **do**
    for all word $v \in V$ **do**
      Calculate $\chi^2(v, a_j)$
    end for
    $T_j = T_j \bigcup \{$ Top ranked n words $\}$
  end for
until No new aspect words are identified or iteration exceeds l
return $T_1, T_2, \ldots$;

Suppose $\mathcal{G}$ is the learnt ranking model parameterized by $\vec{w}_r$, for a new sentence x^i, $\psi(x^i)$ denotes the corresponding bag-of-words features vector, we calculate the ranking score as

$$score_r(x^i) = \mathcal{G}(x^i; \vec{w}_r) = \vec{w}_r \cdot \psi(x^i).$$

Then for all the reviews, our model outputs a ranking score for each sentences.

3.2 Multi-level Sentiment Analysis

For sentiment analysis, we adopt the approach for multi-level sentiment analysis proposed by Yessenalina *et al.*[16]. A benefit of this approach is that it extracts a set of sentences that best represent the polarity of original review only with the supervision of document-level review polarity, which can be easily obtained since many online review websites provide semi-structural reviews with overall ratings. Here, we give a brief description of this approach. A review document is represented by x with corresponding polarity $y \in \{+1, -1\}$. The quality of a sentence with polarity y is computed as

$$q(x^j, y) = \underbrace{y \cdot \vec{w}_{pol}\psi_{pol}(x^j)}_{polarity\ part} + \underbrace{\vec{w}_{subj}\psi_{subj}(x^j)}_{subjective\ part};$$

where $\psi_{pol}(x^j)$ and $\psi_{subj}(x^j)$ denote the polarity and subjectivity features of sentence x^j, $\vec{w}_{pol}$ and $\vec{w}_{subj}$ are learnt weights for polarity and subjectivity features, respectively. It can be seen that the polarity part captures the quality of sentence x^j with polarity y, and the subjective part captures the quality of x^j as a subjective sentence.

Suppose s is a set of sentences that best represents the key opinion of original review x. We define $\mathcal{F}$ parameterized by $\vec{w}_s$ as the function that jointly predicts the document polarity y^* and extracts sentence set s^*, we have

$$(y^*, s^*) = \underset{y \in \{+1,-1\}, s \in \mathcal{P}(x)}{\mathrm{argmax}} \mathcal{F}(x, (y, s); \vec{w}_s)\ ; \qquad (1)$$

where $\mathcal{P}(x)$ is the power set of all the sentences in x. Clearly, $\mathcal{F}$ has the form

$$\mathcal{F}(x,(y,s);\vec{w}_s) = \frac{1}{N(x)}\sum_{x^j\in s} q(x^j,y)$$
$$=\frac{1}{N(x)}\sum_{x^j\in s} y\cdot\vec{w}_{pol}\psi_{pol}(x^j)+\vec{w}_{subj}\psi_{subj}(x^j)\ ;$$

where $N(x)$ is a normalizing factor. As $\psi_{pol}(x^j)$ and $\psi_{subj}(x^j)$ are disjoint by construction, we have

$$\vec{w}_s = [\vec{w}_{pol}, \vec{w}_{subj}];$$

For simplicity, let $\Psi(x,(y,s))$ denote the joint feature map, $\mathcal{F}$ can be written as $\mathcal{F} = \vec{w}_s\Psi(x,(y,s))$. The training process is to optimize the following problem using latent variable structural SVMs [17]:

Optimization Problem 1:

$$\min_{\vec{w},\xi\geq 0}\frac{1}{2}||w||^2 + \frac{C}{N}\sum_{i=1}^{N}\xi_i$$
$$s.t. \forall i:$$
$$\max_{s_i\in\mathcal{P}(x)}\vec{w}_s\Psi(x_i,(y_i,s_i)) \geq \max_{s_i'\in\mathcal{P}(x)}\vec{w}_s\Psi(x_i,(-y_i,s_i'))$$
$$+\Delta(y_i,-y_i,s_i')-\xi_i$$

where C is the regularization parameter, N is the number of training instances. Then we employ the model to jointly predict sentiment and extract a set of sentences that best represent the key polarity of original review using Equation 1.

3.3 Dual Decomposition

Dual decomposition is a general approach for combinatorial optimization, and has been successfully applied to many tasks in natural language processing [13]. For our task here, we expect that sentences in s modeled by multi-level sentiment analysis rank higher in informativeness ranking, i.e., suppose h is the top $|s|$ ranked sentences by the sentence ranking model, our goal is to make an alignment between h and s such that there are as many sentences in common as possible. The joint inference problem is

Joint Inference Problem 1:

$$\underset{(y,s),h}{\text{argmax}}\mathcal{F}(x,(y,s);\vec{w}_s)+\sum_{x^i\in h}\mathcal{G}(x^i;\vec{w}_r)$$
$$s.t.\qquad f(s)=g(h)$$

where f and g are linear functions that map the output s and h to two vectors of length |x|, with 1 for the chosen sentences and 0 elsewhere. To solve the joint inference problem, we introduce a vector of Lagrange multipliers, $\vec{u}\in\mathbb{R}^{|x|}$ to obtain the Lagrangian

$$L((y,s),h,\vec{u}) =\mathcal{F}(x,(y,s);\vec{w}_s)+\sum_{x^i\in h}\mathcal{G}(x^i;\vec{w}_r)$$
$$+\vec{u}\cdot(f(s)-g(h))$$

with the dual objective

$$L(\vec{u}) = \max_{(y,s),h} L((y,s),h,\vec{u}).$$

The optimization can be solved using subgradient algorithm. We initialize the Lagrange multipliers to $\vec{u}^{(0)} = \mathbf{0}$. For $k = 1, 2, \ldots$, and perform the following steps:

$$((y,s)^{(k)},h^{(k)}) = \max_{(y,s),h} L((y,s),h,\vec{u}^{(k-1)}) \qquad (2)$$

followed by

$$\vec{u}^{(k)} = \vec{u}^{(k-1)} - \delta(f(s^{(k)}) - g(h^{(k)}));$$

where δ is the step size. It can be verified that Equation 2 can be solved easily using dual decomposition.

For each review x, we obtain a new ranking model $\mathcal{G}'$ with updated parameters that encoding sentiment information benefitted from dual decomposition. Then we apply the ranking function $\mathcal{G}'$ to rank all the sentences in x, which will naturally make the "sentiment explanations" rank higher.

4. EXPERIMENTS

4.1 Data Preparation

We use the data for explanatory sentence extraction[2][6], which is based on a collection of Amazon product reviews[3] used in[2] and [4]. Kim *et al.*[6] asked 4 labelers to make explanatoriness labels for each sentence with 0 for "no explanation", 1 for "weak explanation" and 2 for "strong explanation". Further, for sentences that are labeled as sentiment explanation, an additional label is introduced with 1 for "less than/equal to half of the text provides good explanation" and 2 for "most of the text provides good explanation". We employ the results of all the labelers, therefore, each sentence has a score ranging from 0 to 16.

Since the test input of our approach is a review, to make evaluations, we ensure that at least one sentence of the review for testing is labeled as sentiment explanation, filtering out those reviews with no sentence labeled as sentiment explanation. Our approach needs training data for the subproblem of multi-level sentiment analysis, we then sample training reviews published from 2004 to 2008 from Amazon product reviews used in [9][4]. For each product domain, we sample 2000 positive (rating greater than or equal to 4) and 2000 negative (rating less than or equal to 2) reviews for training, 500 positive and 500 negative reviews for development. Table 2 presents the statistics of evaluation data[5] where "#." means number of.

domain	camera	cellphone	mp3
#.testing reviews	88	71	137
#.sentiment explanations	87	203	377
#.testing sentences	1,067	814	2,775

Table 2: Data Statistics

4.2 Baselines

To make comparisons, we use the Normalized Discounted Cumulative Gain (nDCG) as the measure to calculate the score of each review, and we choose the following baselines:

- the expected performance of a random ranking, denoted by "*random*";
- SVM^{rank} with only aspect terms, denoted by "*rank(asp)*";
- SVM^{rank} with only sentiment lexicon, denoted by "*rank(op)*";

[2]http://sifaka.cs.uiuc.edu/~hkim277/expSum/
[3]http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html
[4]http://snap.stanford.edu/data/web-Amazon.html
[5]We only use the following product reviews: Canon G3, Nikon coolpix4300, Canon S100, Nokia 6600, Nokia 6600, Creative Labs Nomad Jukebox Zen Xtra 40GB and MicroMP3. For other products, either the reviews are in forms of sentences or the corresponding category can not be easily recognized by product names.

- SVM^{rank} with aspect terms and sentiment lexicon, denoted by "*rank(asp+op)*";
- rule based approach consider aspect terms, sentiment lexicon and sentence length, denoted by "*rule(asp+op)*";

Our approach can be considered as joint inference with *rank(asp+op)* and multi-level sentiment analysis, and we then employ the ranking model with updated parameters of the last iteration in dual decomposition to rank the sentences for a given review.

4.3 Results

Table 3 presents the averaged nDCG score of all the reviews for each product domain. It can be seen that *rank(asp+op)* has a slightly better performance over purely rule based approach *rule (asp+op)*, and our approach achieves a relative high performance compared with baselines.

Domain	camera	cellphone	mp3
random	0.497	0.526	0.484
rank(asp)	0.558	0.662	0.561
rank(op)	0.585	0.609	0.594
rank(asp+op)	0.599	0.669	0.605
rule(asp+op)	0.599	0.644	0.600
Ours	**0.615**	**0.680**	**0.667**

Table 3: Comparison with baselines

5. CONCLUSION AND FUTURE WORK

In this paper, we address the problem of ranking sentiment explanations by formulating the process as two subproblems: sentence informativeness ranking and structural sentiment analysis. Tractable inference is performed through dual decomposition. Preliminary experiments on publicly available data-set demonstrate that our approach is effective and obtains promising performance. For future work, we plan to encode aspect information for fine granular opinion summarization using dual decomposition.

Acknowledgments

This work was partly supported by the following grants from: the National Basic Research Program (973 Program) under grant No. 2012CB316301 & 2013CB329403, the National Science Foundation of China project under grant No. 61332007 and No. 61272227, and the Beijing Higher Education Young Elite Teacher Project.

6. REFERENCES

[1] M. Bonzanini, M. Martinez-Alvarez, and T. Roelleke. Extractive summarisation via sentence removal: Condensing relevant sentences into a short summary. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 893–896, 2013.

[2] X. Ding, B. Liu, and P. S. Yu. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 231–240, 2008.

[3] A. Glaser and H. Schütze. Automatic generation of short informative sentiment summaries. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 276–285, 2012.

[4] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 168–177, 2004.

[5] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 133–142, 2002.

[6] H. D. Kim, M. G. Castellanos, M. Hsu, C. Zhai, U. Dayal, and R. Ghosh. Ranking explanatory sentences for opinion summarization. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 1069–1072, 2013.

[7] T. Lappas, M. Crovella, and E. Terzi. Selecting a characteristic set of reviews. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 832–840, 2012.

[8] T.-Y. Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009.

[9] J. McAuley and J. Leskovec. Hidden factors and hidden topics: Understanding rating dimensions with review text. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, pages 165–172, 2013.

[10] X. Meng and H. Wang. Mining user reviews: From specification to summarization. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, ACLShort '09, pages 177–180, 2009.

[11] B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04, 2004.

[12] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of Empirical methods in natural language processing*, EMNLP '02, pages 79–86, 2002.

[13] A. M. Rush and M. Collins. A tutorial on dual decomposition and lagrangian relaxation for inference in natural language processing. *J. Artif. Int. Res.*, 45(1):305–362, Sept. 2012.

[14] A. M. Rush, D. Sontag, M. Collins, and T. Jaakkola. On dual decomposition and linear programming relaxations for natural language processing. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1–11, 2010.

[15] H. Wang, Y. Lu, and C. Zhai. Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 783–792, 2010.

[16] A. Yessenalina, Y. Yue, and C. Cardie. Multi-level structured models for document-level sentiment classification. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1046–1056, 2010.

[17] C.-N. J. Yu and T. Joachims. Learning structural svms with latent variables. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1169–1176, 2009.

[18] L. Zhuang, F. Jing, and X.-Y. Zhu. Movie review mining and summarization. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, CIKM '06, pages 43–50, 2006.