Expert-Guided Contrastive Opinion Summarization for Controversial Issues

Jinlong Guo[1] Yujie Lu[2] Tatsunori Mori[2] Catherine Blake[1]
[1]Graduate School of Library and Information Science,
University of Illinois at Urban-Champaign, Champaign, IL 61820, USA
[2]Graduate School of Environment and Information Science,
Yokohama National University, Yokohama, 2408501, JAPAN
{jguo24, clblake}@illinois.edu, {luyujie, mori}@forest.eis.ynu.ac.jp

ABSTRACT
This paper presents a new model for the task of contrastive opinion summarization (COS) particularly for controversial issues. Traditional COS methods, which mainly rely on sentence similarity measures are not sufficient for a complex controversial issue. We therefore propose an Expert-Guided Contrastive Opinion Summarization (ECOS) model. Compared to previous methods, our model can (1) integrate expert opinions with ordinary opinions from social media and (2) better align the contrastive arguments under the guidance of expert prior opinion. We create a new data set about a complex social issue with "sufficient" controversy and experimental results on this data show that the proposed model are effective for (1) producing better arguments summary in understanding a controversial issue and (2) generating contrastive sentence pairs.

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; Contrastive opinion summarization; Topic model; Controversial issue; Similarity

1. INTRODUCTION
With the continuing growth of information on the web and especially online social media (e.g. Twitter, Facebook, etc.), there is a great amount of opinionated text created every day, which stimulates the research area of opinion mining (sentiment analysis). Though there's much progress in opinion mining techniques in recent years, particularly for product review, finding out contrastive arguments that are for (pros) or against (cons) a controversial issue is still a challenging task.

Extraction of comparative sentences with contrasting opinion is a recently proposed problem in [6]. In the paper, the authors proposed the new task as contrastive opinion summarization (COS). For COS, the opinion labels of the sentences (positive vs. negative) are given beforehand, and the system aims at selecting the most representative and comparable sentences accordingly.

However, as traditional opinion mining research is mainly motivated by business (e.g. product review, movie review), the proposed model is mainly tested on product reviews. While the model is useful for digesting reviews, we argue that it's not sufficient for mining opinion about complex controversial issues. The latter task is more challenging because, unlike product review mining where the aspects/features of a product are usually limited and explicit, the aspects/arguments for a controversial issue are much more complicated, nuanced and abstract. For a controversial issue, it's helpful to display as many opinions as possible and in a contrastive way. Study has shown that people are willing to read more when presented with arguments from both sides [17].

In this study, we propose an Expert-Guided Contrastive Opinion Summarization (ECOS) model that could leverage a few expert opinions in mining huge amount of ordinary opinions (e.g. opinions from Twitter). Our model could (1) integrate expert prior opinions and ordinary opinions from social media, thus providing a comprehensive picture of the controversial issue under debate and (2) output contrastive argument pairs for better understanding the "controversy". We will adopt a semi-supervised PLSA model for argument clustering, which is the key part of the ECOS model. (Note that we assume one document, in our case one tweet, belongs to one argument category, thus topic numbers represent argument numbers) The rationale is twofold: first, the semi-supervised PLSA model can easily integrate expert arguments (as prior) with ordinary arguments; second, since both argument sides (positive and negative) are guided by expert arguments, it's easier to output contrastive argument pairs.

Our model can be beneficial for a diverse group of users like politicians, campaign leaders, and policy makers to make sense of what happened around a controversial topic for informed decisions. The contributions of this paper are as follows: (1) we proposed a new model (ECOS) for contrastive opinion summarization (COS) for complex controversial issues, which can integrate expert opinions and ordinary opinions and output contrastive argument

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2743038

pairs. (2) We create an annotated data set with "sufficient" controversy for mining complex controversial issues. (3) We run experiments to test the proposed methods on our data set and show that our methods are effective.

The rest of the paper is organized as follows. In Section 2, we discuss related work. In Section 3, we explain the proposed model: Expert-Guided Contrastive Opinion Summarization (ECOS). In Section 4 and 5 we detail the two key steps in ECOS, namely the semi-supervised PLSA model and sentence selection strategy. In section 6 we describe our dataset and experiment settings. Section 7 reports the results and discussion of our experiment. We give the conclusion in Section 8.

2. RELATED WORK

In this section, we will briefly discuss some related works to our contrastive opinion summarization task.

The first line of work is the general area of opinion summarization. There are a lot of research done in this area [5, 3], and in recent years, models that integrate different sources for opinion summarization have drawn much attention. For example, [8, 9] propose to use expert and ontology to organize opinions. However, their models do not deal with controversial opinions.

The second line of work that is highly related to ours is opinion mining on Twitter. Some works adapt traditional topic modeling methods on Twitter mining [4, 10]. There are also models that deal with topic and sentiment simultaneously for mining Twitter opinion [7]. In terms of controversy mining, we are aware of two types of research. The first is to detect controversial events in Twitter. These works aim at capturing events/topics that are assumed to be controversial [13, 14]. Such models only predict whether tweets/events are controversial or not without dealing with what's the controversy (namely the arguments for the controversial issue). The second type of research is to predict user's attitude towards controversial issues [2]. This is also different from our goal of summarizing controversial arguments.

The third line of research that is related to our work is perspective modeling. These works are related to variants of the original topic models and aim to model different perspective/viewpoint of the text collection besides the topics [11, 1]. These models can capture some contrastive perspectives/viewpoints for certain topic, but not to provide a comprehensive summarization of arguments for controversial issues.

The last line of research that are closely related to ours is contrastive opinion summarization. The task was first proposed in [6], and most tested for product reviews. Recent works try to improve the methods proposed in the original paper based on the same product review data [16]. [18] is highly similar to our work, in which they provide a three-step framework to retrieve perspectives for controversial issues. However, the proposed method is not well tested. [12] presents a two-stage approach to summarizing multiple contrastive viewpoints in opinionated text. However, with only unsupervised methods, it's unlikely to extract the actual viewpoints or arguments that the users really want for a controversial issue, which has the same limitation in [11, 1].

3. EXPERT-GUIDED CONTRASTIVE OPINION SUMMARIZATION

Traditional opinion summarization aims to select a set of the most representative opinionated text from an input set. While contrastive opinion summarization (COS) aims to capture not only the most representative sentences but also the ones that have contrastive meaning. This type of opinion summarization can help user better digest different opinions on different arguments of an issue under debate.

In this paper, we propose the Expert-Guided Contrastive Opinion Summarization (ECOS) model that integrates expert opinions with ordinary opinions and produces a list of representative and contrastive argument pairs for the controversial issue. For example, in the case of "gay marriage", people would like to see arguments that is for and against "gay marriage" with regard to religion. It would be helpful that the system can return a pair of sentences/tweets, such as

Positive side (pros): "*@stupotwakefield hypocrisy in church etc church not moving with times/gay marriage…*"

Negative side (cons): "*Where are all the riots in support of gay marriage? That's right, you "Christians" pick and choose which part of the bible you support.*"

There are a few motivations to integrate expert opinions in the COS model. First, unlike product review, we find that opinions for complex controversial (social) issues are difficult to model without any prior knowledge. Without the guidance of the expert prior, the argument clustering process could be very arbitrary, leading to meaningless clustering results. Second, in order to make informed decisions, it's necessary to obtain as diverse opinions (both expert and ordinary opinions) as possible for controversial issues. Finally, from the perspective of COS model, our expert guided model can help to solve the problem of alignment between argument clusters in positive and negative sentence sets. With the emergence of websites that provide edited expert opinions for different controversial topics (e.g. procon.org; debate.org), we can acquire expert opinions easily nowadays.

In previous COS methods, similarity between sentences with the same sentiment labels (content similarity) is used to find out the most representative sentences and similarity between sentences with different sentiment labels (contrastive similarity) is used to find out the most contrastive sentences. The COS task therefore is defined as an optimization problem where the goal is to maximize the representativeness and contrastiveness of the input sentence set. Different methods are proposed based on this framework, for example Representativeness First and Contrastiveness First strategy in [6], Contrastive Max-Sum Opinion Summarization in [16].

However, arguments space is complicated for a controversial issue, a pure unsupervised approach [6, 16] is not sufficient. Instead of defining the COS task as an optimization problem, we adopt a more heuristic approach with the idea of adding expert prior opinions. First, we cluster arguments under the guidance of expert prior opinions for both positive and negative (sentence/tweet) sets. We then select representative sentence(s) from each aligned cluster as contrastive argument pairs. For unaligned clusters, we further use the similarity-based approaches to select contrastive sentences. Specifically we propose a semi-supervised PLSA model to cluster arguments. We will detail the semi-supervised PLSA model and sentence selection strategy in the following two sections.

4. SEMI-SUPERVISED PLSA MODEL

Topic models have been widely used in different types of text mining tasks recently. In [8], the authors show that a semi-supervised PLSA model can be used to integrate expert and ordinary opinions and generate useful aligned integrated opinion

summaries. We will adopt this model to cluster the arguments under the guidance of expert prior information.

The basic PLSA model can extract topics/arguments from our opinion sentences/tweets. The semi-supervised PLSA model can extract topics/arguments that "mapped" with the expert opinions. In probabilistic models, this can be achieved by extending the basic PLSA to incorporate a conjugate prior defined based on the expert opinion segments and using the Maximum A Posterior (MAP) estimator instead of the Maximum Likelihood estimator as used in basic PLSA. The MAP estimate can be computed using essentially the same EM algorithm as in the basic PLSA model with only slightly different updating formula for the component language models. The new updating formula is:

$$\mathrm{p}(\mathrm{w}|\theta_j) = \frac{\sum_{d\epsilon C} c(w,d)p(z_{d,w,j}) + \sigma_j p(w|r_j)}{\sum_{w'\in V}\sum_{d'\in C} c(w',d')p(z_{d',w',j}) + \sigma_j} \quad (1)$$

In this formula, $p(w|r_j)$ denotes the probability of words in expert opinion segments (suppose there are j expert argument categories). σ_j denotes the confidence parameter for the expert prior. In the original paper, $p(w|r_j)$ is calculated using formula (2), where the probability of words in a specific expert opinion segment is the count of words in that segment divided by count of all words in that expert segment. Heuristically only aspect/topic words (such as nouns) are calculated and opinion words (such as adjectives, adverbs and verbs) are removed.

$$\mathrm{p}(\mathrm{w}|r_i) = \frac{c(w,r_i)}{\sum_{w'\epsilon V} c(w',r_i)} \quad (2)$$

In our approach, we propose two ways to calculate $p(w|r_j)$. One way is the same as formula (2), where the expert opinion segments can be obtained from websites like procon.org. The other way is to manually assign prior probability to keywords by the user according to his/her understanding of the controversial issue under exploration (this is particularly useful for explorative analysis).

After estimating the parameters for the semi-supervised topic model, we could group each document (in our case, sentence/tweet) into one of the topic/argument cluster by choosing the topic model with the largest probability of generating the document using formula (3):

$$\arg\max p(d_i|\theta_j) = \arg max \sum_{w\in V} c(w,d_i)p(w|\theta_j) \quad (3)$$

For a more detailed description of the semi-supervised PLSA model, please refer to [8].

5. STRATEGY FOR SENTENCE SELECTION

Since there's no guarantee that the ordinary opinions can be well mapped with expert opinions, (e.g. Twitter users may publish new ideas that are not included in the expert prior), the next key step is to select sentences from different clusters. We first define the key measure used in our method, namely contrastive similarity and then describe our sentence selection strategy.

The contrastive similarity is meant to measure how well two sentences with opposite opinions match up with each other. In [6] the authors adopt the heuristics of removing sentiment related words (e.g. negation words and adjectives) in calculating contrastive similarity. We will experiment two versions of similarity calculation, one with all words, and the other with only content words. In terms of sentence similarity measure, we adopt the common cosine similarity measure.

Our sentence selection strategy has two parts: those aligned clusters (this means both clusters in positive set and negative set are mapped with expert prior category, we call this aligned cluster) and those unaligned clusters (if either cluster in two sides is not mapped with expert prior category).

For aligned clusters, we calculate the contrastive similarity for every two sentence pair from the cluster and chose sentence pair with the highest contrastive similarity measure. Note that we use expert prior keywords to see whether a cluster is mapped with the expert opinion.

For one-side mapped clusters, which means only one side (either positive or negative set) is mapped with expert prior category, we delete the mapped cluster and leave the unmapped cluster (of sentences) as candidate for further consideration. The rationale here is if only one side is mapped with expert prior category while the other is not, it is highly likely that the other side does not cover this argument/topic. In such case, we would delete the cluster of the mapped side since it is unlikely that the other side has aligned sentences. We keep the unmapped cluster for further consideration because since it's not mapped with expert category, it could possibly be any category.

For clusters not mapped with expert opinion in either side, which means the category of the cluster is not clear, we would keep both clusters as candidate for further sentence selection.

Finally for all the "free" candidate clusters from above steps, we can adopt different kinds of sentence selection strategies. In fact, at this point, our problem has come back to the initial stage of a COS task, with a bunch of sentences from both the positive set and negative set. As a consequence, theoretically we can apply any COS methods reported in previous literature. Since our model mainly depends on the semi-supervised part (namely the aligned clusters), we only propose two simple strategies here for sentence selection from "free" candidate clusters. The two strategies comprise the two baselines in the following experiment section.

The first strategy (Baseline 1) is only contrastive similarity based. In this strategy, we discard the cluster information and select top k pairs of sentences with the highest contrastive similarity. This is a simplified version of the Contrastiveness First strategy in [6] that mainly focus on contrastiveness instead of representativeness of the sentences.

The second strategy (Baseline 2) considers cluster information, assuming that cluster information is useful to express representativeness. The strategy we use for sentence pair selection is: (1) select top m (m=2 in our case) sentences from each cluster with the highest probability (result output from previous topic modeling); (2) for each cluster in the positive (negative) set, calculate the contrastive similarity with all sentences in the negative (positive) set; (3) choose sentence pairs that have similarity measure more than n (n=0.15 in our case) or if no sentence pair satisfies the threshold, then choose the sentence pair with highest similarity as candidate sentence pairs. The intuition is that if we only choose the sentence pair with highest probability for each cluster, we might miss some pairs with high contrastive similarity that are very likely to be the correct pairs; (4) rank the candidate sentence pairs according to contrastive similarity and choose the top k (k=10 in our case) sentence pairs.

6. EXPERIMENTS

6.1 Data Collecting

Since there's no annotated data set for complex controversial issue with "sufficient" controversy, we decided to create our own data set. We chose the "gay marriage" case to test our model for several reasons. First, this case provides "sufficient" controversy with diverse arguments for opinion mining. Second, we can get expert opinion data from an expert website (procon.org) for this case. We can also easily get enough Twitter data about this case that represent ordinary opinions.

For this experiment, we collected expert opinion from this website under the topic "gay marriage"[1]. The website is updated regularly. By 2015-1-21, there are 15 arguments for pros and 13 arguments for cons on the "gay marriage" page. We also fetched data from Twitter using Twitter API with the keyword "gay marriage". We collected more than two-week Twitter data from 2014-11-19 to 2014-12-05. We used a subset of one-day data with the largest amount of tweets (7624 tweets) as experiment data in this paper.

6.2 Data Preprocessing

As for expert opinions got from the procon.org website, we further convert the 15 arguments for pros and 13 arguments for cons into 8 argument themes (some of the expert pros and cons are about the same argument themes) and identify some keywords for each argument themes. These keywords will be employed as prior in our semi-supervised PLSA model.

As for Twitter data, which contains rich features and more verbal expressions, we use a Twitter specific preprocessing tool named TweetNLP [15] for feature tagging. Besides regular POS tags like noun, verb, etc., the tool can identify a set of tweet-specific tags like hashtag, mention, URL, emoticon, etc. with a reported accuracy of 93%.

We then filter out tweets that are not useful as informative arguments by removing tweets with URL, retweets and tweets that are too short. We got 633 out of 7624 tweets that we believe contain more useful information about opinions/arguments for understanding of the controversial issue.

6.3 Data Annotations & Gold Standard Summary

In order to get the gold standard summary, we ask a scholar in LGBT (Lesbian, Gay, Bisexual and Transgender) studies to annotate the data for us. In the first stage, the annotation label is positive, negative, neutral and none of the above. In the second stage, we randomly select 50 positive and 50 negative sentences and ask our expert to annotate/cluster the sentences (here means tweets) according to the reasons/arguments about the issue. If tweets in different sides get the same argument category label, it is assumed as correct pairs. The result of the annotation is summarized in Table 1.

As a result, there are 7 clusters in the positive set, among which 5 clusters are within our expert prior opinion categories, and 8 clusters in the negative set, among which 5 are within the expert prior categories. The extra two categories in the positive set is "Emotions" and "Liberalism" and the extra three categories in the negative set is "Emotions", "Liberalism" and "Priority". We do not count the "other" category in both sets (which means no specific reason for or against gay marriage), since we only care about arguments. This result shows different coverage of topics/arguments between expert and Twitter opinions and gives us a sense of the difficulty of the task (e.g. some of the clusters only contain 1 sentence, which makes it very difficult to output the correct argument pair).

Table 1. Gold Standard Tweets for Argument Clustering

Topic ID	Expert Prior Category	Positive	Negative
Topic 1	**Children & Adoption**	None	3
Topic 2	**Economic Problem**	None	None
Topic 3	**Civil Rights**	9	1
Topic 4	**Discrimination**	3	1
Topic 5	**Tradition & Definition**	2	4
Topic 6	**Psychological Problem**	None	None
Topic 7	**Religion**	6	12
Topic 8	**Procreation**	2	None
Topic 9	Emotions	6	1
Topic 10	Liberalism	2	4
Topic 11	Priority	None	6
	Other	20	18

6.4 Experiment Settings and Parameter Tuning

Our experiment is based on the annotated data that includes 50 positive tweets and 50 negative tweets, with an average word token for each tweet 21.92 and 20.34 respectively. Slightly different preprocessing is used for different stages in our model. For the input of semi-supervised PLSA model, we remove all the Twitter specific features and keep only word tokens. We also use NLTK for stop word removal and stemming (Porter stemmer). For tweet contrastive similarity measure, we keep two versions of sentence representation. One is the whole word token, the other is only content word token.

For the semi-supervised PLSA model, we set expert prior σ as 100 (strong prior). We also set σ as 0 as a baseline to test the clustering results without expert prior opinion. We will run the model 10 times and use the results with the highest data likelihood. In terms of the number of clusters k, since we have 8 categories of prior expert opinions, we heuristically set k=10 to map the expert opinions as well as capture new opinions.

7. RESULTS & DISCUSSION

7.1 Example results and qualitative analysis

Table 2 shows some example results of our model. We can see some interesting results returned by the system. (1) If only one side of the cluster is mapped to expert prior category, we can conjecture that the corresponding argument is not mentioned in the other side. This somewhat reveals which argument is more often used to support (or oppose) an issue. For example, we can see that "bad environment for children" is more used as arguments against gay marriage (Topic 1); while "procreation is

1. http://gaymarriage.procon.org/

not the only purpose of marriage" is more often used to argue for gay marriage (Topic 8). (2) If both sides of the cluster are mapped to expert prior category (aligned cluster), we can see that tweets can supplement the expert opinions. For example, in Topic 5, tweets contribute arguments about the discussion of the definition of marriage "You need to look up what marriage is, because gay marriage is not redefining it, considering the history." (3) If neither side of the cluster is mapped with the expert prior category, we could conjecture that this particular argument is not well discussed in Twitter. Topic 2 shows the example where expert arguments about economic problems concerning gay marriage are not covered in Twitter. (4) For those unaligned clusters, our system can also output useful new argument pairs (beyond expert argument pairs). Topic 9 produces the pair of argument about "liberalism" where it is not covered in our expert prior cate gory. (Refer to Table 1 when reading this topic result)

Table 2. Sample Opinion Results from Twitter

Topic ID	Tweet Positive	Tweet Negative
Topic 1	None	@jimmyjazz68 I'm guessing that the opposition is more likely about "harm❓? to children in gay marriage, not sole purpose of breeding.
Topic 2	None	None
Topic 5	**@TheFullBug You need to look up what marriage is, because gay marriage is not redefining it, considering the history.**	**@IngrahamAngle absolutely. Same with feminist movement, gay marriage, etc. really just to undermine the institutions that exist**
Topic 8	if those against allowing gay marriage base their position on breeding, then let us also disallow marriage between heterosexual non-breeders	None
Topic 9	**extreme conservatives love to tout out the slippery slope that gay marriage would bring were it put in place.the one we're on is far scarier**	**@AnaKasparian @SairoHusky damn slippery slopes, just like the gays, first you legalize gay marriage, then we all start practicing bestiality**

7.2 Baseline and quantitative analysis

We adopt the measures proposed in [6] to evaluate our model quantitatively. **Precision**: The precision of a summary with k contrastive pairs is defined as the percentage of the k pairs that are agreed by our expert annotator. If a retrieved pair exists in the evaluator's (expert) paired-cluster set, it is assumed that the pair is "relevant". Thus precision is basically the number of such agreed pairs divided by k. Precision tells us how contrastive the sentence pairs of the summary are. **Coverage**: The coverage of a summary is the percentage of aligned clusters (gold standard cluster pair) covered in the summary. If a pair of sentences appears in an aligned pair of clusters, it is assumed that the aligned cluster is covered. Coverage measures the argument representativeness of a summary.

We run two baseline algorithms to compare our model. Baseline 1 uses only the contrastive similarity measure between positive and negative sentences. We simply choose the top k (k=10) sentence pairs with the highest contrastive similarity. This baseline aims to see how only similarity measure performs for the task. Baseline 2 adds the cluster information, which uses the "free" clustering without the expert prior information (see section 5 for detailed strategy).

Table 3 shows the result of baseline algorithms and our ECOS model. Compared to Baseline 1 that does not use any clustering information, strategy that uses clustering could improve coverage, though both baselines do not perform well, suggesting that a simple sentence similarity measure is not sufficient for a complex controversial issue reported here.

Our expert guided model shows significant improvement in both precision (0.600) and coverage (0.667). Results on our data set show that 4 clusters are aligned, and a simple sentence strategy of choosing the highest contrastive similarity among the two aligned clusters can achieve 100% accuracy. For unaligned clusters, after deleting one-side mapped clusters, we have 5 "free" clusters in the positive set and 5 "free" clusters in the negative set. We got 2 out of 6 sentence pairs from the "free" clusters. We therefore get a final precision of 0.600 and coverage of 0.667. Given our data is much more complicated than the product review data in [6], this result shows the effectiveness of our semi-supervised model for contrastive opinion summarization.

We also compare contrastive similarity based on different sentence representation. Contrary to previous research, contrastive similarity based on all word tokens perform better. We postulate that this is probably because our model dose not mainly depends on the simple similarity measure.

Table 3. Measures for Baselines and ECOS Model

	Word All		Content Word Only	
	Precision	**Coverage**	**Precision**	**Coverage**
Baseline 1	0.200	0.167	0.100	0.167
Baseline 2	0.200	0.333	0.200	0.333
ECOS	0.600	0.667	0.400	0.667

7.3 Discussion

We can see that if tweets are clustered in the right expert-guided cluster, then there is high probability that the sentence pair chosen from the paired clusters can be correct. In our experiment, the 4 sentence pairs chosen from the 4 aligned clusters are all correct. While only 2 of the sentence pairs out of 6 chosen from the "free" candidate clusters are correct.

We can improve our model in two directions. On one hand, the clustering result can be improved. An error analysis of the result reveals at least two types of error. The first is due to word sense ambiguity. For example we have the expert prior keyword "right" for the argument category "civil rights". However, tweets containing the term "right" with different meaning will be wrongly clustered. This type of error can be reduced by word disambiguation. Another type of error of clustering is when more expert prior words appear in one tweet. Actually we found that a hard clustering that clusters a sentence into only one cluster might not be optimal for controversial issue because some arguments or

topics are related (a tweet might belong to two related argument types). A soft clustering strategy might be useful.

One the other hand, the contrastive similarity plays an important role in sentence pair selection. Our intuition is that for complex controversial issues, more advanced semantic based sentence similarity measure might be helpful.

Finally we need to point out that the COS output of k most representative sentence pairs might not be the optimal output for complex controversial issues, thus the measurement of Precision and Coverage adopted here is also not optimal. For each aligned cluster, only outputting one optimal sentence pair is not the most helpful way in understanding the "complexity". As we have shown in the qualitative analysis, more sentence pairs output from Twitter under the expert argument category contribute to the understanding of the issue. We suggest better opinion/argument output structure for controversial issue.

8. CONCLUSION

Traditional opinion summarization does not output contrasting sentences for comparative summarization. In order to deal with the problem, the contrastive opinion summarization (COS) problem has been introduced. Based on previous work on COS, we proposed a new model for COS particularly for complex controversial issue (ECOS), which can integrate expert opinion with ordinary opinion and output contrastive sentence pairs. We created our own data set for testing our model. The results show that: (1) Our model provides the potential of integrating expert and ordinary opinions (both positive and negative) in a unified way for users to better digest opinions concerning controversial issue. (2) Compared with previous COS methods, our expert guided COS model proves effective with regard to precision and coverage.

9. ACKNOWLEDGMENTS

We thank Xueqing Liu for helping implementing the PLSA model and the anonymous reviewers for their useful comments.

10. REFERENCES

[1] Fang, Y., Si, L., Somasundaram, N., & Yu, Z. (2012, February). Mining contrastive opinions on political texts using cross-perspective topic model. In *Proceedings of the fifth ACM international conference on Web search and data mining* (pp. 63-72). ACM.

[2] Gao, H., Mahmud, J., Chen, J., Nichols, J., & Zhou, M. (2014). Modeling User Attitude toward Controversial Topics in Online Social Media. In *the Eighth International AAAI Conference on Weblogs and Social Media (ICWSM 2014)*.

[3] Ganesan, K., Zhai, C., & Viegas, E. (2012, April). Micropinion generation: an unsupervised approach to generating ultra-concise summaries of opinions. In *Proceedings of the 21st international conference on World Wide Web* (pp.869-878). ACM.

[4] Hong, L., & Davison, B. D. (2010, July). Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics* (pp. 80-88). ACM.

[5] Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 168-177). ACM.

[6] Kim, H. D., & Zhai, C. (2009, November). Generating comparative summaries of contradictory opinions in text. In *Proceedings of the 18th ACM conference on Information and knowledge management* (pp. 385-394). ACM.

[7] Lim, K. W., & Buntine, W. (2014, November). Twitter Opinion Topic Model: Extracting Product Opinions from Tweets by Leveraging Hashtags and Sentiment Lexicon. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* (pp. 1319-1328). ACM.

[8] Lu, Y., & Zhai, C. (2008, April). Opinion integration through semi-supervised topic modeling. In *Proceedings of the 17th international conference on World Wide Web* (pp. 121-130). ACM.

[9] Lu, Y., Duan, H., Wang, H., & Zhai, C. (2010, August). Exploiting structured ontology to organize scattered online opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 734-742). Association for Computational Linguistics.

[10] Mehrotra, R., Sanner, S., Buntine, W., & Xie, L. (2013, July). Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval* (pp. 889-892). ACM.

[11] Paul, M., & Girju, R. (2010). A two-dimensional topic-aspect model for discovering multi-faceted topics. *Urbana*, *51*, 61801.

[12] Paul, M. J., Zhai, C., & Girju, R. (2010, October). Summarizing contrastive viewpoints in opinionated text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 66-76). Association for Computational Linguistics.

[13] Popescu, A. M., & Pennacchiotti, M. (2010, October). Detecting controversial events from twitter. In *Proceedings of the 19th ACM international conference on Information and knowledge management* (pp. 1873-1876). ACM.

[14] Popescu, A. M., Pennacchiotti, M., & Paranjpe, D. (2011, March). Extracting events and event descriptions from twitter. In *Proceedings of the 20th international conference companion on World Wide Web* (pp. 105-106). ACM.

[15] Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., & Smith, N. A. (2013). Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters. In *HLT-NAACL* (pp. 380-390).

[16] Özsoy, M. G., & Cakıcı, R. (2014). Contrastive Max-Sum Opinion Summarization. In *Information Retrieval Technology* (pp. 256-267). Springer International Publishing.

[17] Vydiswaran, V. G., Zhai, C., Roth, D., & Pirolli, P. (2012, October). BiasTrust: Teaching biased users about controversial topics. In *Proceedings of the 21st ACM international conference on Information and knowledge management* (pp. 1905-1909). ACM.

[18] Zheng, W., & Fang, H. (2010). A Retrieval System based on Sentiment Analysis. *HCIR 2010*, 52.