Exploiting Structured Ontology to Organize Scattered Online Opinions

Yue Lu, Huizhong Duan, Hongning Wang, ChengXiang Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign
{yuelu2,duan9,wang296,czhai}@illinois.edu

Abstract

We study the problem of integrating scattered online opinions. For this purpose, we propose to exploit structured ontology to obtain well-formed relevant aspects to a topic and use them to organize scattered opinions to generate a structured summary. Particularly, we focus on two main challenges in implementing this idea, (1) how to select the most useful aspects from a large number of aspects in the ontology and (2) how to order the selected aspects to optimize the readability of the structured summary. We propose and explore several methods for solving these challenges. Experimental results on two different data sets (US Presidents and Digital Cameras) show that the proposed methods are effective for selecting aspects that can represent the major opinions and for generating coherent ordering of aspects.

1 Introduction

The explosive growth of online opinions raises interesting challenges for opinion integration and summarization. It is especially interesting to integrate and summarize scattered opinions in blog articles and forums as they tend to represent the general opinions of a large number of people and get refreshed quickly as people dynamically generate new content, making them valuable for understanding the current views of a topic.

However, opinions in blogs and forums are usually fragmental, scattered around, and buried among other off-topic content, so it is quite challenging to organize them in a meaningful way. Traditional text summarization techniques generate an unstructured list of sentences as a summary, which cannot reveal representative opinions on different aspects of a topic or effectively facilitate navigation into the huge opinion space. To address this limitation, recent work has shown the usefulness of generating a structured summary of opinions, in which related opinions are grouped into topical aspects with explicit labeling of all the aspects. A major challenge in producing such a structured summary is how to generate these aspects for an arbitrary topic (e.g., products, political figures, policies, etc.). Intuitively, the aspects should be concise phrases that can both be easily interpreted in the context of the topic under consideration and capture the major opinions. However, where can we find such phrases and which phrases should we select as aspects? Furthermore, once we selected aspects, how should we order them to improve the readability of a structured summary? One way to generate aspects is to cluster all the opinion sentences and then identify representative phrases in each cluster. Although aspects selected in this way can effectively capture the major opinions, a major limitation is that it is generally hard to ensure that the selected phrases are well connected with the given topic (Chen and Dumais, 2000).

In this paper, we propose a novel approach to generating aspects by leveraging the ontologies with structured information that are available online, such as open domain knowledge base in Freebase[1]. Such kind of ontology data is not in small scale by any measure. For example, Freebase alone contains more than 10 million topics, 3000 types, and 30,000 properties; moreover, it is constantly growing as people collaboratively contribute. Freebase provides different properties for different types of topics such as personal information for a "US President" and product features for a "Digital Camera". Since this kind of resources can provide related entities/relations for a

---

[1] http://www.freebase.com

wide range of topics , our general idea is to leverage them as guidance for more informed organization of scattered online opinions, and in particular, to select the most important properties of a topic from such structured ontology as aspects to generate a structured opinion summary. A significant advantage of this approach to aspect generation is that the selected aspects are guaranteed to be very well connected with the topic, but it also raises an additional challenge in selecting the aspects to best capture the major opinions from a large number of aspects provided for each topic in the ontology. Different from some existing work on exploiting ontologies, e.g., (Sauper and Barzilay, 2009), which relies on training data, we focus on exploring *unsupervised* approaches, which can be applied to a larger scope of topics.

Specifically, given a topic with entries in an ontology and a collection of scattered online opinions about the topic, our goal is to generate a structured summary where representative major opinions are organized with well aligned aspects and in an order easy for human to follow. We propose the following general approach: First, retrieval techniques are employed to align opinions to relevant aspects. Second, a subset of most interesting aspects are selected. Third, we will further order the selected aspects to present them in a reasonable order. Finally, for the opinions uncovered by the selected aspects from the ontology, we use a phrase ranking method to suggest new aspects to add to the ontology for increasing its coverage.

Implementing the second and third steps involves special challenges. In particular, without any training data, it is unclear how we should show the most interesting aspects in ontology with major opinions aligned and which presentation order of aspects is natural and intuitive for human. Solving these two challenges is the main focus of this paper. We propose three methods for aspect selection, i.e., size-based, opinion coverage-based, and conditional entropy-based methods, and two methods for aspect ordering, i.e., ontology-ordering and coherence ordering. We evaluate our methods on two different types of topics: US Presidents and Digital Cameras. Qualitative results demonstrate the utility of integrating opinions based on structured ontology as well as

the generalizability of proposed methods. Quantitative evaluation is also conducted to show the effectiveness of our methods.

Note that we use the term "opinion" to broadly refer to any discussion in opinionated sources such as blogs and reviews. This allows us to formulate and solve the problem in a general way. Indeed, the main goal of our work is to extract and organize the major opinions about a topic that are buried in many scattered opinionated sources rather than perform deeper understanding of opinions (e.g., distinguishing positive from negative opinions), which can be done by using any existing sentiment analysis technique as an orthogonal post-processing step after applying our method.

2 Related Work

Aspect summarization, i.e., structured opinion summarization over topical aspects, has attracted much attention recently. Existing work identifies aspects using frequent-pattern/association-rule mining, e.g. (Liu et al., 2005; Popescu and Etzioni, 2005), sentence clustering, e.g. (Gamon et al., 2005; Leouski and Croft, 1996), or topic modeling, e.g. (Mei et al., 2006; Titov and McDonald, 2008). After that, meaningful and prominent phrases need to be selected to represent the aspects, e.g. (Zhao and He, 2006; Mei et al., 2007). However, these methods suffer from the problem of producing trivial aspects. Consequently, some of the aspects generated are very difficult to interpret (Chen and Dumais, 2000). In this paper, we propose a different kind of approach that is to use aspects provided by ontology which are known to be relevant and easy to interpret.

Ontology is used in (Carenini et al., 2005) but only for mapping product features. The closest work to ours are (Lu and Zhai, 2008; Sauper and Barzilay, 2009); both try to use well-written articles for summarization. However, (Lu and Zhai, 2008) assumes the well-written article is structured with explicit or implicit aspect information, which does not always hold in practice, while (Sauper and Barzilay, 2009) needs a relatively large amount of training data in the given domain. In comparison, our work only needs the ontology information for the given topic which is much easier to obtain from resources such as Freebase.

3 Methods

Given (1) an input topic T, (2) a large number of aspects/properties $A = \{A_1, ..., A_m\}$ from an ontology that are related to T, and (3) a huge collection of scattered opinion sentences about the topic $D_T = \{s_1, \ldots, s_n\}$, our goal is to generate a structured organization of opinions that are both aligned well with the interesting aspects and representative of major opinions about the topic.

The envisioned structured organization consists of a sequence of selected aspects from ontology ordered to optimize readability and a set of sentences matching each selected aspect. Once we obtain a set of sentences in each aspect, we can easily apply a standard text summarization method to further summarize these sentences, thus the unique challenges related to our main idea of exploiting ontology are the following, which are also the main focus of our study:

Aspect Selection: How can we select a subset of aspects $A' \subset A$ to capture the *major* opinions in our opinion set D_T?

Aspect Ordering: How can we order a subset of selected aspects A' so as to present them in an order $\pi(A')$ that is most natural with respect to human perception?

New Aspects Suggestion: Can we exploit the opinions in D_T to suggest new aspects to be added to the ontology?

3.1 Aspect Selection

In order to align the scattered opinions to the most relevant aspects, we first use each aspect label $A_i \in A$ as a query to retrieve a set of relevant opinions in the collection $S_i \subseteq D_T$ with a standard language modeling approach, i.e., the KL-divergence retrieval model (Zhai and Lafferty, 2001). Up to 1000 opinion sentences are retrieved for each aspect; each opinion sentence can be potentially aligned to several aspects. In this way, scattered online discussion are linked to the most relevant aspects in the ontology, which enables a user to use aspects as "semantic bridges" to navigate into the opinion space..

However, there are usually a lot of candidate aspects in an ontology, and only some are heavily commented in online discussions, so showing all the aspects is not only unnecessary, but also overwhelming for users. To solve this problem,

we propose to utilize the aligned opinions to further select a subset of the most interesting aspects $A' \subset A$ with size k. Several approaches are possible for this subset selection problem.

Size-based: Intuitively, the selected subset A' should reflect the major opinions. So a straightforward method is to order the aspects A_i by the size of the aligned opinion sentences S_i, i.e., the number of relevant opinion sentences, and then select the top k ones.

Opinion Coverage-based: The previous method does not consider possible redundancy among the aspects. A better approach is to select the subset that covers as many *distinct* opinion sentences as possible. This can be formulated as a maximum coverage problem, for which a greedy algorithm is known to be a good approximation: we select one aspect at a time that is aligned with the largest number of uncovered sentences.

Conditional Entropy-based: Aspects from a structured ontology are generally quite meaningful, but they are not designed specifically for organizing the opinions in our data set. Thus, they do not necessarily correspond well to the natural clusters in scattered opinions. To obtain aspects that are aligned well with the natural clusters in scattered opinions, we can first cluster D_T into l clusters $C = \{C_1, \ldots, C_l\}$ using K-means with $TF \times IDF$ as features, and then choose the subset of aspects that minimize Conditional Entropy of the cluster label given the aspect:

$$
\begin{aligned}
A' &= \arg\min H(C|A') = \arg\min \\
&\left[-\sum_{A_i \in A', C_i \in C} p(A_i, C_i) \log \frac{p(A_i, C_i)}{p(A_i)} \right]
\end{aligned}
$$

This Conditional Entropy measures the uncertainty about the cluster label of a sentence given the knowledge of its aspect. Intuitively, if the aspects are aligned well with the clusters, we would be able to predict well the cluster label of a sentence if we know its aspect, thus there would be less uncertainty about the cluster label. In the extreme case when the cluster label can be completely determined by the aspect, the conditional entropy would reach its minimum (i.e., 0). Intuitively, the conditional entropy-based method essentially selects the most appropriate aspects from

Algorithm 1 Greedy Algorithm for Conditional Entropy Based Aspect Selection

Input: $A = \{A_1, ..., A_m\}$
Output: k-sized $A' \subseteq A$
1:   $A' = \{\cup_{i=1}^{m} A_i\}$
2:   **for** j=1 to k **do**
3:     $bestH = \infty$; $bestA = A_0$
4:     **for** each A_i in A **do**
5:       $tempA' = \{A_i, A' \setminus A_i\}$
6:       **if** $H(C|tempA') < bestH$ **then**
7:         $bestH = H(C|tempA')$
8:         $bestA = A_i$
9:     $A' = \{bestA, A' \setminus bestA\}$
10: output A'

the ontology to label clusters of opinions.

The exact solution of this combinatorial optimization problem is NP-complete, so we employ a polynomial time greedy algorithm to approximate it: in the i-th iteration, we select the aspect that can minimize the conditional entropy given the previous $i - 1$ selected aspects. Pseudo code is given in Algorithm 1.

3.2 Aspect Ordering

In order to present the selected aspects to users in a most natural way, it is important to obtain a coherent order of them, i.e., generating an order consistent with human perception. To achieve this goal, our idea is to use human written articles on the topic to learn how to organize the aspects automatically. Specifically, we would order aspects so that the relative order of the sentences in all the aspects would be as consistent with their order in the original online discussions as possible.

Formally, the input is a subset of selected aspects A'; each $A_i \in A'$ is aligned with a set of relevant opinion sentences $S_i = \{S_{i,1}, S_{i,2}, ...\}$. We define a coherence measurement function over sentence pairs $Co(S_{i,k}, S_{j,l})$, which is set to 1 iff $S_{i,k}$ appears before $S_{j,l}$ in the same article. Otherwise, it is set to 0. Then a coherence measurement function over an aspect pair can be calculated as

$$Co(A_i, A_j) = \frac{\sum_{S_{i,k} \in S_i, S_{j,l} \in S_j} Co(S_{i,k}, S_{j,l})}{|S_i||S_j|}$$

As an output, we would like to find a permutation $\hat{\pi}(A')$ that maximizes the coherence of all pairwise aspects, i.e.,

$$\hat{\pi}(A') = \arg\max_{\pi(A')} \sum_{A_i, A_j \in A', A_i \prec A_j} Co(A_i, A_j)$$

Algorithm 2 Greedy Algorithm for Coherence Based Aspect Ordering

Input: A
Output: $\pi(A)$
1:   **for** each A_i, A_j in A **do**
2:     calculate $Co(A_i, A_j)$
3:   **for** $p = 1$ to $len = A.size()$ **do**
4:     $Max = A[1]$
5:     **for** each aspect A_i in A **do**
6:       $A_i.coherence = 0$
7:       **for** each aspect A_j in $\pi(A)$ **do**
8:         $A_i.coherence{+} = Co(A_j, A_i)$
9:       **for** each aspect A_j in A, $j \neq i$ **do**
10:       $A_i.coherence{+} = Co(A_i, A_j)$
11:      **if** $A_i.coherence > Max.coherence$ **then**
12:       $Max = A_i$
13:     remove Max from A; add Max to $\pi(A)$
14: output $\pi(A)$

where $A_i \prec A_j$ means that A_i is before A_j. It is easy to prove that the problem is NP-complete. Therefore, we resort to greedy algorithms to find approximations of the solution. Particularly we view the problem as a ranking problem. The algorithm proceeds by finding at each ranking position an aspect that can maximize the coherence measurement, starting from the top of the rank list. The detailed algorithm is given in Algorithm 2.

3.3 New Aspects Suggestion

Finally, if the opinions cover more aspects than in the ontology, we also want to identify informative phrases to label such extra aspects; such phrases can also be used to further augment the ontology with new aspects.

This problem is similar to existing work on generating labels for clusters (Zeng et al., 2004) or topic models (Mei et al., 2007). Here we employ a simple but representative technique to demonstrate the feasibility of discovering interesting new aspects for augmenting the ontology. We first extract named entities from scattered opinions D_T using Stanford Named Entity Recognizer (Finkel et al., 2005). After that, we rank the phrases by pointwise Mutual Information (MI):

$$MI(T, ph) = \log \frac{P(T, ph)}{P(T)P(ph)}$$

where T is the given topic and ph refers to a candidate entity phrase. $P(T, ph)$ is proportional to the number of opinion sentences they co-occur; $P(T)$ or $P(ph)$ are proportional to the number of times T or ph appears. A higher MI value indicates a

Statistics	Category 1 US president	Category 2 Digital Camera
Number of Topics	36	110
Number of Aspects	65±26	32±4
Number of Opinions	1001±1542	170±249

Table 1: Statistics of Data Sets

stronger association. We can then suggest the top ranked entity phrases that are not in the selected aspects as new aspects.

4 Experiments

4.1 Data Sets

To examine the generalizability of our methods, we test on two very different categories of topics: *US Presidents* and *Digital Cameras*.[2] For the ontology, we leverage Freebase, downloading the structured ontology for each topic. For the opinion corpus, we use blog data for US Presidents and customer reviews for Digital Cameras. The blog entries for US Presidents were collected by using Google Blog Search[3] with the name of a president as the query. Customer reviews for Digital Cameras were crawled from CNET[4]. The basic statistics of our data sets is shown in Table 1. For all the data collections, Porter stemmer (Porter, 1997) is applied and stop words are removed.

4.2 Sample Results

We first show sample results of automatic organization of online opinions. We use the opinion coverage-based algorithm to select 10 aspects (10-20 aspects were found to be optimal in (Käki, 2005)) and then apply the coherence-based aspect ordering method. The number of clusters is set so that there are on average 15 opinions per cluster.

Opinion Organization: Table 2 and Table 3 present sample results for President Ronald Reagan and Sony Cybershot DSC-W200 camera respectively[5]. We can see that (1) although Freebase aspects provide objective and accurate information about the given topics, extracted opinion sentences offer additional subjective information; (2) aligning scattered opinion sentences to most relevant aspects in the ontology helps digestion and

---

navigation; and (3) the support number, which is the number of opinion sentences aligned to an aspect, can show the popularity of the aspect in the online discussions.

Adaptability of Aspect Selection: Being unsupervised is a significant advantage of our methods over most existing work. It provides flexibility of applying the methods in different domains without the requirement of training data, benefiting from both the ontology based template guidance as well as data-driven approaches. As a result, we can generate different results for different topics even in the same domain. In Table 4, we show the top three selected and ordered aspects for Abraham Lincoln and Richard Nixon. Although they belong to the same category, different aspects are picked up due to the differences in online opinions. People talk a lot about Lincoln's role in American Civil War and his famous quotation, but when talking about Nixon, people focus on ending the Vietnam war and the Watergate scandal. "Date of birth" and "Government position" are ranked first because people tend to start talking from these aspects, which is more natural than starting from aspects like "Place of death".

Baseline Comparison: We also show below the aspects for Lincoln generated by a representative approach using clustering method (e.g. (Gamon et al., 2005)). i.e., we label the largest clusters by selecting phrases with top mutual information. We can see that although some phrases make sense, not all are well connected with the given topic; using aspects in ontology circumvents this problem. This example confirms the finding in previous work that the popular existing clustering-based approach to aspects generation cannot generate meaningful labels (Chen and Dumais, 2000).

```
Vincent
New Salem State Historic Site
USS Abraham Lincoln
Martin Luther King Jr
Gettysburg
John F.
```

New Aspect Discovery: Finally, in Table 5 we show some phrases ranked among top 10 using the method described in Section 3.3. They reveal additional aspects covered in online discussions and serve as candidate new aspects to be added to Freebase. Interestingly, John Wilkes Booth, who assassinated President Lincoln, is not explicitly

FreeBase Aspects	Supt	Representative Opinion Sentences
Appointees: - Martin Feldstein - Chief Economic Advisor	897	Martin Feldstein, whose criticism of Reagan era deficits has not been forgotten. Reagan's first National Security advisor was quoted as declaring...
Government Positions Held: - President of the United States - Jan 20, 1981 to Jan 20, 1989	967	1981 Jan 20, Ronald Reagan was sworn in as president as 52 American hostages boarded a plane in Tehran and headed toward freedom. 40th president of the US Ronald Reagan broke the so called "20 year curse"...
Vice president: - George H. W. Bush	847	8 years, 1981-1988 George H. W. Bush as vice president under Ronald Reagan... ...exception to the rule was in 1976, when George H W Bush beat Ronald.

Table 2: Opinion Organization Result for President Ronald Reagan

FreeBase Aspects	Supt	Representative Opinion Sentences
Format: - Compact	13	Quality pictures in a compact package. ... amazing is that this is such a small and compact unit but packs so much power.
Supported Storage Types: - Memory Stick Duo	11	This camera can use Memory Stick Pro Duo up to 8 GB Using a universal storage card and cable (c'mon Sony)
Sensor type: - CCD	10	I think the larger ccd makes a difference. but remember this is a small CCD in a compact point-and-shoot.
Digital zoom: -2×	47	once the digital :smart" zoom kicks in you get another 3x of zoom I would like a higher optical zoom, the W200 does a great digital zoom translation...

Table 3: Opinion Organization Result for Sony Cybershot DSC-W200 Camera

listed in Freebase, but we can find it in people's online discussion using mutual information.

4.3 Evaluation of Aspect Selection

Measures: Aspect selection is a new challenge, so there is no standard way to evaluate it. It is also very hard for human to read all of the aspects and opinions and then select a gold standard subset. Therefore, we opt to use indirect measures capturing different characteristics of the aspect selection problem (1) *Aspect Coverage* (AC): we first assign each aspect A_i to the cluster C_j that has the most overlapping sentences with A_i, approximating the cluster that would come into mind when a reader sees A_i. Then AC is defined as the percentage of the clusters covered by at least one aspect. (2) *Aspect Precision* (AP): for each covered cluster C_i, AP measures the Jaccard similarity between C_i as a set of opinions and the union of all aspects assigned to C_i. (3) *Average Aspect Precision* (AAP): defines averaged AP for all clusters where an uncovered C_i has a zero AP; it essentially combines AC and AP. We also report *Sentence Coverage* (SC), i.e., how many distinct opinion sentences can be covered by the selected aspects and *Conditional Entropy* (H), i.e., how well the selected aspects align with the natural clusters in the opinions; a smaller H value indicates a better alignment.

Results: We summarize the evaluation results in

Measures	SC	H	AC	AP	AAP
PRESIDENTS					
Random	503	1.9069	0.5140	0.0933	0.1223
Size-based	500	1.9656	0.3108	**0.1508**	0.0949
Opin Cover	**746**	1.8852	0.5463	0.0913	0.1316
Cond Ent.	479	**1.7687**	**0.5770**	0.0856	**0.1552**
CAMERAS					
Random	55	1.6389	0.6554	0.0871	0.1271
Size-based	70	1.6463	0.6071	**0.1077**	0.1340
Opin Cover	**82**	1.5866	0.6998	0.0914	0.1564
Cond Ent.	70	**1.5598**	**0.7497**	0.0789	**0.1574**

Table 6: Evaluation Results for Aspect Selection

Table 6. In addition to the three methods described in Section 3.1, we also include one baseline of averaging 10 runs of random selection. The best performance by each measure on each data set is highlighted in bold font. Not surprisingly, opinion coverage-based approach has the best sentence coverage (SC) performance and conditional entropy-based greedy algorithm achieves the lowest H. Size-based approach is best in aspect precision but at the cost of lowest aspect coverage. The trade-off between AP and AC is comparable to that between precision and recall as in information retrieval while AAP summarizes the combination of these two. The greedy algorithm based on conditional entropy outperforms all other approaches in AC and also in AAP, suggesting that it can provide a good balance between AP and AC.

Supt	Richard-Nixon	Supt	Abraham-Lincoln
50	Date of birth: - Jan 9, 1913	419	Government Positions Held: - United States Representative Mar 4,1847-Mar 3,1849
108	Tracks Recorded: - 23-73 Broadcast: End of the Vietnam War	558	Military Commands: - American Civil War - United States of America
120	Works Written About This Topic: - Watergate	810	Quotations: - Nearly all men can stand adversity, but if you want to test a man's character, give him power.

Table 4: Comparison of Aspect Selection for Two Presidents (aligned opinions are omitted here)

Suggested Phrases	Supporting Opinion Sentences
Abraham Lincoln Presidential Library	CDB projects include the Abraham Lincoln Presidential Library and Museum
Abraham Lincoln Memorial	..., eventually arriving at Abraham Lincoln Memorial.
John Wilkes Booth	John Wilkes Booth shoots President Abraham Lincoln at Ford's Theatre ...

Table 5: New Phrases for Abraham Lincoln

4.4 Evaluation of Aspect Ordering

Human Annotation: In order to quantitatively evaluate the effectiveness of aspect ordering, we conduct user studies to establish gold standard ordering. Three users were each given k selected aspects and asked to perform two tasks for each US President: (1) identify clusters of aspects that are more natural to be presented together (*cluster constraints*) and (2) identify aspect pairs where one aspect is preferred to appear before the other from the viewpoint of readability. (*order constraints*). We did not ask them to provide a full order of the k aspects, because we suspect that there are usually more than one "perfect" order. Instead, identifying partial orders or constraints is easier for human to perform, thus provides more robust gold standard.

Human Agreement: After obtaining the human annotation results, we first study human consensus on the ordering task. For both types of human identified constraints, we convert them into pairwise relations of aspects, e.g., "A_i and A_j should be presented together" or "A_i should be displayed before A_j". Then we calculate the agreement percentage among the three users. In Table 7, we can see that only a very small percentage of pair-wise partial orders (15.92% of the cluster constraints and none of the order constraints) are agreed by all the three users, though the agreement of clustering is much higher than that of ordering. This indicates that ordering the aspects is a subjective and difficult task.

Measures: Given the human generated gold standard of partial constraints, we use the following measures to evaluate the automatically gen-

AgreedBy	Cluster Constraint	Order Constraint
1	37.14%	89.22%
2	46.95%	10.78%
3	15.92%	0.00%

Table 7: Human Agreement on Ordering

erated full ordering of aspects: (1) *Cluster Precision* (pr_c): for all the aspect pairs placed in the same cluster by human, we calculate the percentage of them that are also placed together in the system output. (2) *Cluster Penalty* (p_c): for each aspect pair placed in the same cluster by human, we give a linear penalty proportional to the number of aspects in between the pair that the system places; p_c can be interpreted as the average number of aspects between aspect pairs that should be presented together in the case of misordering. Smaller penalty corresponds to better ordering performance. (3) *Order Precision* (pr_o): the percentage of correctly predicted aspect pairs compared with human specified order.

Results: In Table 8, we report the ordering performance based on two selection algorithms: opinion coverage-based and conditional entropy-based. Different selection algorithms provide different subsets of aspects for the ordering algorithms to operate on. For comparison with our coherence-based ordering algorithm, we include a random baseline and Freebase ontology ordering. Note that Freebase order is a very strong baseline because it is edited by human even though the purpose was not for organizing opinions. To take into account the variation of human annotation, we use four versions of gold standard: three are from the individual annotators and one from the union of their annotation. We did not include the gold stan-

Selection Algo	Gold STD	Cluster Random	Precision Freebase	(pr_c) Coherence	Cluster Random	Penalty Freebase	(p_c) Coherence	Order Random	Precision Freebase	(pr_o) Coherence
Opin Cover	1	0.3290	**0.9547**	0.9505	1.8798	0.1547	**0.1068**	0.4804	**0.7059**	0.4510
Opin Cover	2	0.3266	**0.9293**	0.8838	1.7944	0.3283	**0.1818**	**0.4600**	0.4000	0.4000
Opin Cover	3	0.2038	**0.4550**	0.4417	2.5208	**1.3628**	1.7994	0.5202	0.4561	**0.5263**
Opin Cover	union	0.3234	**0.7859**	0.7237	1.8378	0.6346	**0.4609**	**0.4678**	0.4635	0.4526
Cond Entropy	1	0.2540	**0.9355**	0.8978	2.0656	0.2957	**0.2016**	0.5106	**0.7111**	0.5444
Cond Entropy	2	0.2535	0.7758	**0.8323**	2.1790	0.7530	**0.5222**	0.4759	**0.6759**	0.5093
Cond Entropy	3	0.2523	0.4030	**0.5545**	2.3079	2.1328	**1.1611**	0.5294	0.7143	**0.8175**
Cond Entropy	union	0.3067	0.7268	**0.7488**	1.9735	1.0720	**0.7196**	0.5006	0.6500	**0.6833**

Table 8: Evaluation Results on Aspect Ordering

dard that is the intersection of three annotators because that would leave us with too little overlap.

We have several observations: (1) In general, results show large variations when using different versions of gold standard, indicating the subjective nature of the ordering task. (2) Coherence-based ordering shows similar performance to Freebase order-based in cluster precision (pr_c), but when we take into consideration the distance-based penalty (p_c) of separating aspects pairs in the same cluster, coherence-based ordering is almost always significantly better except in one case. This shows that our method can effectively learn the coherence of aspects based on how their aligned opinion sentences are presented in online discussions. (3) Order precision (pr_o) can hardly distinguish different ordering algorithm. This indicates that people vary a lot in their preferences as which aspects should be presented first. However, in cases when the random baseline outperforms others the margin is fairly small, while Freebase order and coherence-based order have a much larger margin of improvement when showing superior performance.

5 Conclusions and Future Work

A major challenge in automatic integration of scattered online opinions is how to organize all the diverse opinions in a meaningful way for any given topic. In this paper, we propose to solve this challenge by exploiting related aspects in structured ontology which are guaranteed to be meaningful and well connected to the topic. We proposed three different methods for selecting a subset of aspects from the ontology that can best capture the major opinions, including size-based, opinion coverage-based, and conditional entropy-based methods. We also explored two ways to order aspects, i.e., ontology-order and coherence optimization. In addition, we also proposed appropriate measures for quantitative evaluation of both aspect selection and ordering.

Experimental evaluation on two data sets (US President and Digital Cameras) shows that by exploiting structured ontology, we can generate interesting aspects to organize scattered opinions. The conditional entropy method is shown to be most effective for aspect selection, and the coherence optimization method is more effective than ontology-order in optimizing the coherence of the aspect ordering, though ontology-order also appears to perform reasonably well. In addition, by extracting salient phrases from the major opinions that cannot be covered well by any aspect in an existing ontology, we can also discover interesting new aspects to extend the existing ontology.

Complementary with most existing summarization work, this work proposes a new direction of using structured information to organize and summarize unstructured opinions, opening up many interesting future research directions. For instance, in order to focus on studying aspect selection and ordering, we have not tried to optimize sentences matching with aspects in the ontology; it would be very interesting to further study how to accurately retrieve sentences matching each aspect. Another promising future work is to organize opinions using both structured ontology information and well-written overview articles.

References

Carenini, Giuseppe, Raymond T. Ng, and Ed Zwart. 2005. Extracting knowledge from evaluative text. In *K-CAP '05: Proceedings of the 3rd international conference on Knowledge capture*, pages 11–18, New York, NY, USA. ACM.

Chen, Hao and Susan Dumais. 2000. Bringing order to the web: automatically categorizing search results. In *CHI '00: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 145–152, New York, NY, USA. ACM.

Finkel, Jenny Rose, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370, Morristown, NJ, USA. Association for Computational Linguistics.

Gamon, Michael, Anthony Aue, Simon Corston-Oliver, and Eric K. Ringger. 2005. Pulse: Mining customer opinions from free text. In Famili, A. Fazel, Joost N. Kok, José María Peña, Arno Siebes, and A. J. Feelders, editors, *IDA*, volume 3646 of *Lecture Notes in Computer Science*, pages 121–132. Springer.

Käki, Mika. 2005. Optimizing the number of search result categories. In *CHI '05: CHI '05 extended abstracts on Human factors in computing systems*, pages 1517–1520, New York, NY, USA. ACM.

Leouski, Anton V. and W. Bruce Croft. 1996. An evaluation of techniques for clustering search results. Technical report.

Liu, Bing, Minqing Hu, and Junsheng Cheng. 2005. Opinion observer: analyzing and comparing opinions on the web. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pages 342–351, New York, NY, USA. ACM.

Lu, Yue and Chengxiang Zhai. 2008. Opinion integration through semi-supervised topic modeling. In Huai, Jinpeng, Robin Chen, Hsiao-Wuen Hon, Yunhao Liu, Wei-Ying Ma, Andrew Tomkins, and Xiaodong Zhang, editors, *WWW*, pages 121–130. ACM.

Mei, Qiaozhu, Chao Liu, Hang Su, and ChengXiang Zhai. 2006. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 533–542.

Mei, Qiaozhu, Xuehua Shen, and ChengXiang Zhai. 2007. Automatic labeling of multinomial topic models. In Berkhin, Pavel, Rich Caruana, and Xindong Wu, editors, *KDD*, pages 490–499. ACM.

Pang, Bo and Lillian Lee. 2007. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Popescu, Ana-Maria and Oren Etzioni. 2005. Extracting product features and opinions from reviews. In *HLT '05*, pages 339–346, Morristown, NJ, USA. Association for Computational Linguistics.

Porter, M. F. 1997. An algorithm for suffix stripping. pages 313–316.

Sauper, Christina and Regina Barzilay. 2009. Automatically generating wikipedia articles: A structure-aware approach. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 208–216, Suntec, Singapore, August. Association for Computational Linguistics.

Titov, Ivan and Ryan McDonald. 2008. Modeling online reviews with multi-grain topic models. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 111–120, New York, NY, USA. ACM.

Zeng, Hua-Jun, Qi-Cai He, Zheng Chen, Wei-Ying Ma, and Jinwen Ma. 2004. Learning to cluster web search results. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 210–217, New York, NY, USA. ACM.

Zhai, Chengxiang and John Lafferty. 2001. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of CIKM 2001*, pages 403–410.

Zhao, Jing and Jing He. 2006. Learning to generate labels for organizing search results from a domain-specified corpus. In *WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 390–396, Washington, DC, USA. IEEE Computer Society.