Comprehensive Review of Opinion Summarization

Hyun Duk Kim
University of Illinois at Urbana-Champaign
Kavita Ganesan
University of Illinois at Urbana-Champaign
Parikshit Sondhi
University of Illinois at Urbana-Champaign
and
ChengXiang Zhai
University of Illinois at Urbana-Champaign

General Terms: Text Processing, Survey

Additional Key Words and Phrases: Opinion summarization, Opinion Mining

1. INTRODUCTION

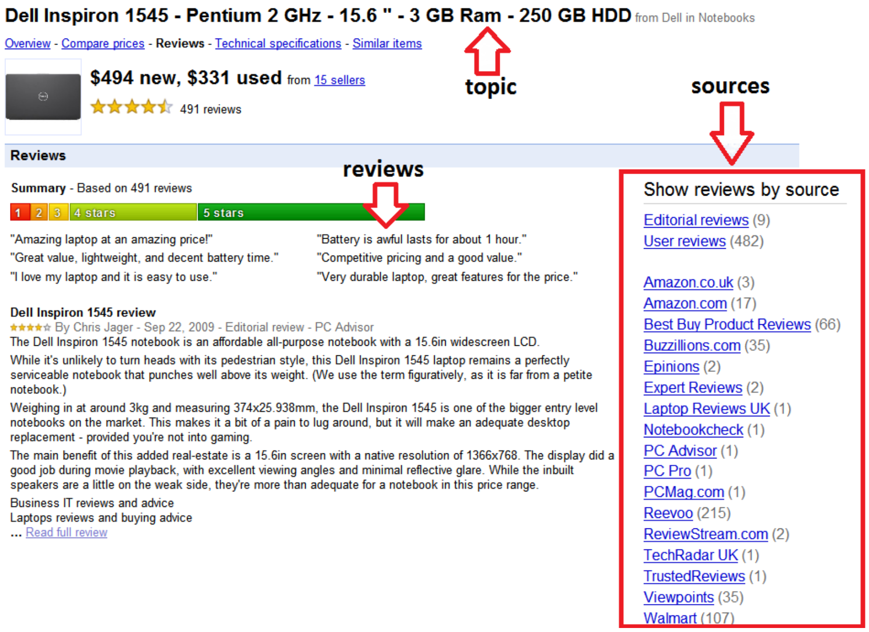
Last updated on: 2013/08/10

With the growth of the web over the last decade, opinions can now be found almost everywhere - blogs, social networking sites like Facebook and Twitter, news portals, e-commerce sites, etc. While these opinions are meant to be helpful, the vast availability of such opinions becomes overwhelming to users when there is just too much to digest. Consider a user looking to buy a laptop. Figure 1 shows all the available laptop reviews for a *Dell laptop* obtained from Google Product Search. Although these opinions are meant for just one product, there are more than 400 reviews for this one product from around 20 different sources. Such overwhelming amounts of information make summarization of the web very critical.

Over the last few years, this special task of summarizing opinions has stirred tremendous interest amongst the *Natural Language Processing* (NLP) and *Text Mining* communities. 'Opinions' mainly include opinionated text data such as blog/review articles, and associated numerical data like aspect rating is also included. While different groups have different notions of what an opinion summary should be, we consider any study that attempts to generate a concise and digestible summary of a large number of opinions as the study of *Opinion Summarization*.

The simplest form of an opinion summary is the result of sentiment prediction (by aggregating the sentiment scores). The task of sentiment prediction or classification itself has been studied for many years. Beyond such summaries, the newer generation of opinion summaries includes structured summaries that provide a well-organized breakdown by aspects/topics, various formats of textual summaries and temporal visualization. The

Authors' address: Hyun Duk Kim (hkim277@illinois.edu), Kavita Ganesan (kganes2@illinois.edu), Parikshit Sondhi (sondhi1@illinois.edu) and ChengXiang Zhai (czhai@cs.uiuc.edu), University of Illinois at Urbana-Champaign, 201 N Goodwin Ave Urbana, IL 61801



Dell Inspiron 1545 - Pentium 2 GHz - 15.6 " - 3 GB Ram - 250 GB HDD from Dell in Notebooks

Overview - Compare prices - **Reviews** - Technical specifications - Similar items

$494 new, $331 used from 15 sellers

★★★★☆ 491 reviews

Reviews

Summary - Based on 491 reviews

"Amazing laptop at an amazing price!"
"Great value, lightweight, and decent battery time."
"I love my laptop and it is easy to use."

"Battery is awful lasts for about 1 hour."
"Competitive pricing and a good value."
"Very durable laptop, great features for the price."

Dell Inspiron 1545 review
★★★★☆ By Chris Jager - Sep 22, 2009 - Editorial review - PC Advisor
The Dell Inspiron 1545 notebook is an affordable all-purpose notebook with a 15.6in widescreen LCD.

While it's unlikely to turn heads with its pedestrian style, this Dell Inspiron 1545 laptop remains a perfectly serviceable notebook that punches well above its weight. (We use the term figuratively, as it is far from a petite notebook.)

Weighing in at around 3kg and measuring 374x25.938mm, the Dell Inspiron 1545 is one of the bigger entry level notebooks on the market. This makes it a bit of a pain to lug around, but it will make an adequate desktop replacement - provided you're not into gaming.

The main benefit of this added real-estate is a 15.6in screen with a native resolution of 1366x768. The display did a good job during movie playback, with excellent viewing angles and minimal reflective glare. While the inbuilt speakers are a little on the weak side, they're more than adequate for a notebook in this price range.

Business IT reviews and advice
Laptops reviews and buying advice
... Read full review

Show reviews by source

Editorial reviews (9)
User reviews (482)

Amazon.co.uk (3)
Amazon.com (17)
Best Buy Product Reviews (66)
Buzzillions.com (35)
Epinions (2)
Expert Reviews (2)
Laptop Reviews UK (1)
Notebookcheck (1)
PC Advisor (1)
PC Pro (1)
PCMag.com (1)
Reevoo (215)
ReviewStream.com (2)
TechRadar UK (1)
TrustedReviews (1)
Viewpoints (35)
Walmart (107)

Fig. 1. Example Google product search research on Dell laptop [1]

different formats of summaries complement one another by providing a different level of understanding. For example, sentiment prediction on reviews of a product can give a very general notion of what the users feel about the product. If the user needs more specifics, then the topic-based summaries or textual summaries may be more useful. Regardless of the summary formats, the goal of opinion summarization is to help users digest the vast availability of opinions in an easy manner. The approaches utilized to address this summarization task vary greatly and touch different areas of research including text clustering, sentiment prediction, text mining, NLP analysis, and so on. Some of these approaches rely on simple heuristics, while others use robust statistical models.

Currently, there are three surveys that are related to the study of opinion summarization. Chapter 11 of Liu's book [Liu 2006] covers various techniques in opinion mining and summarization. In the book, Liu first defines the notion of 'opinion' and 'opinion mining' and introduces basic concepts related to these definitions. Then he describes techniques in opinion mining covering sentiment classification, opinion summarization, and opinion spam detection. While Liu summarizes this area with a novel framework, as this book was published in 2006, his survey does not cover some of the more recent work in opinion summarization. A big portion of Liu's book is dedicated to explaining definitions and techniques in sentiment classification (the simplest form of an opinion summary), and only a small portion of his book discusses the task of summary generation beyond sentiment classification. In addition, most of the opinion summarization works discussed by Liu are rule-based and heuristics-oriented techniques, missing out on some of the probabilistic methods that were published during the same time period.

In 2010, Liu wrote another book chapter about 'Sentiment Analysis and Subjectivity'

[Liu 2010]. Although the new book chapter covers some recent articles, the content in general is very similar to the previous book chapter. The focus of the new book chapter is purely sentiment classification techniques, not covering some of the state-of-the-art opinion summarization methods. As there are already multiple surveys touching the sentiment classification task, in our survey, we focus purely on the recent techniques used in opinion summarization that goes beyond sentiment classification or uses sentiment classification as one of the components in summarization.

Pang and Lee's survey [Pang and Lee 2008] on Opinion Mining and Sentiment Analysis provides a better coverage of works related to opinion summarization. Although this survey covers a lot of recent works, it is focused on 'opinion mining' broadly rather than opinion 'summarization'. In Pang's survey, the methods are explained at a very high level, and the classification of related works is different from the view we will take. In Pang's survey, works in opinion summarization are categorized as single document, multi-documents, textual and visual approaches. In our survey, we will provide a breakdown of the techniques used into distinct steps (e.g. step1: aspect/feature extraction, step 2: sentiment prediction, and step 3: summary generation) and attempt to classify the techniques used in each step to provide both a broad perspective and detailed understanding of those techniques. By focusing on the smaller scope of study, we are able to use more sophisticated categorization for opinion summarization. This will allow readers to compare and contrast the approaches with ease.

In this survey, we cover a comprehensive list of state-of-the-art techniques and paradigms used for the task of opinion summarization that goes beyond sentiment classification. We will classify the approaches in various ways and describe the techniques used in an intuitive manner. We will also provide various aspects of evaluation in opinion summarization, which was not covered by other previous surveys. Finally, we will provide insights into the weaknesses of the approaches and describe the challenges that remain to be solved in this area.

The rest of the paper is organized as follows. We first introduce the related topics in Section 2 where we provide background information of the related research areas involved in opinion summarization. Then, we go on to describing the representative opinion summarization approaches, aspect-based opinion summarization, and techniques used for it. After that, we discuss non-aspect-based approaches in Section 5. In Section 6, we discuss various aspects of evaluation of opinion summarization techniques. In Section 7, we conclude with a discussion on the open challenges that remain to be solved.

2. BACKGROUND

Before we describe various opinion summarization works, we will provide background knowledge on some important relevant topics used to solve the problem of opinion summarization. Many works in opinion summarization are built upon some closely related research areas such as sentiment classification, text summarization, etc. In this section, we will briefly discuss some of the core related areas used to build opinion summarization systems. Many of these topics are covered by Pang's survey [Pang and Lee 2008].

2.1 Sentiment Classification

Sentiment classification focuses on determining the semantic orientation of words, sentences, and documents. The earliest works in sentiment classification were at the level of individual words. The usual approach is to extract adjectives from texts and try to identify their orientation. Different approaches were proposed for the purpose later. [Hatzivassiloglou and McKeown 1997] utilized the linguistic constraints on semantic orientations of adjectives. [Kamps and Marx 2001] proposed a WordNet-based approach, using semantic distance from a word to "good" and "bad" as a classification criterion. [Turney 2002] used pointwise mutual information (PMI) as semantic distance between two words to measure sentiment strength of words. Later, [Turney and Littman 2003] found that using cosine distance in latent semantic analysis (LSA) space as a distance measure leads to better accuracy.

Among the works of the document level classification, the earliest work was done by [Pang et al. 2002] who experimented with several machine learning techniques with common text features to classify movie reviews. Authors presented a number of further refinements in their subsequent works [Pang and Lee 2004; 2005]. Another good evaluation for various sentiment classification methods based on reviews was given by [Dave et al. 2003]. They experimented with a number of methods for designing sentiment classifiers using training corpus. Other related works in this regard includes [Osgood et al. 1967; Wilson et al. 2004; Mullen and Collier 2004].

Sentiment classification has been used in opinion summarization as one of the most important key steps. While the results of sentiment classification can be used as a simple summary in itself, the notion of opinion summarization involves much more than just identifying orientations of phrases, sentences or documents. Opinion summarization approaches provide a holistic method starting from some raw opinionated text up to the generation of human understandable summaries.

2.2 Subjectivity Classification

Subjectivity classification aims at differentiating sentences, paragraphs, or documents that present opinions/evaluations from those that present factual information. [Wiebe 2000] attempted to find high quality adjective features by word clustering. [Riloff and Wiebe 2003; Riloff et al. 2003] used subjective nouns learned automatically from un-annotated data. [Yu and Hatzivassiloglou 2003] presented a Bayesian approach to identify if a document is subjective or not.

Subjectivity classification, however, is different from sentiment classification in that the former only aims at finding if an opinion is present or not and does not attempt to identify the orientation of these opinions. Sometimes subjectivity classification is used as an input data preprocessing step for sentiment classification. By filtering out objective sentences in advance of sentiment classification, subjectivity classification can increase the accuracy of sentiment classification.

2.3 Text Summarization

There are two representative types of automatic summarization methods. *Extractive Summary* is a summary made by selecting representative text segments, usually sentences, from the original documents. *Abstractive Summary* does not use the existing sentences from the input data; it analyzes documents and directly generates sentences. Because it is hard to generate readable and complete sentences, studies on extractive summary are more popular than that on abstractive summary.

Research in the area of summarizing documents focused on proposing paradigms for extracting salient sentences from text and coherently organizing them to build a summary of the entire text. The relevant works in this regard includes [Paice 1990; Kupiec et al.

1995; Hovy and Lin 1999]. While the earlier works focused on summarizing a single document, later, researchers started to focus on summarizing multiple documents.

Due to the characteristics of data itself, opinion summarization has different aspects from the classic text summarization problem. In an opinion summary, usually the polarities of input opinions are crucial. Sometimes, those opinions are provided with additional information such as rating scores. Also, the summary formats proposed by the majority of the opinion summarization literature are more structured in nature with the segmentation by topics and polarities.

However, text summarization techniques still can be useful in opinion summarization when text selection and generation step. After separating input data by polarities and topics, classic text summarization can be used to find/generate the most representative text snippet from each category.

2.4 Topic Modeling

Topic model is a generative probabilistic model which uses vocabulary distribution to find topics of texts. Topic modeling captures word frequencies and co-occurrences effectively. For example, if word A and B co-occur regularly, word A and C never co-occur, we can assume there is one topic including word A and B, and there is a different topic including C. Representative topic modeling approaches are Probabilistic Latent Semantic Analysis (PLSA) [Hofmann 1999] and Latent Dirichlet Analysis (LDA) [Blei et al. 2003].

The goal of topic modeling is to identify a set of topics or themes from a large collection of documents. Based on topic probability, researchers try to identify documents that are relevant to each of themes. For example, in a document collection being comprised of laptop reviews, some of the themes may be battery life, cost, warranty, etc. It is clear that many of these themes represent product features around which opinions may need to be summarized. If words used in positive documents are very different from those used in negative documents, topic modeling may identify positive topics and negative topics. Thus, topic modeling approaches can be greatly useful in automatic identification of features as well as sentiment classification for opinion summarization. Topic modeling naturally normalizes features as clusters, and users do not need to worry about complicated parameter tuning. Also, if there is existing knowledge to incorporate, we can use prior probabilities. Depending on prior knowledge of topics, each topic can work as a feature or one of sentiment orientations.

3. CONCEPTUAL FRAMEWORK FOR OPINION SUMMARIZATION

Going by our scope of survey and definition of opinion summarization, the related body of work in this area can be very broadly classified into those that require a set of aspects and those that do not rely on the presence of aspects. We can call them as *aspect-based summarization* and *non-aspect-based summarization*. Figure 3 shows the overview of our opinion summarization classification.

Aspect-based summarization divides input texts into aspects, which are also called as features and subtopics, and generates summaries of each aspect. For example, for the summary of 'iPod', there can be aspects such as 'battery life', 'design', 'price', etc. By further segmenting the input texts into smaller units, aspect-based summarization can show more details in a structured way. Aspect segmentation can be even more useful when overall opinions are different from opinions of each aspect because aspect-based summary can present opinion distribution of each aspect separately. The aspect-based approaches are

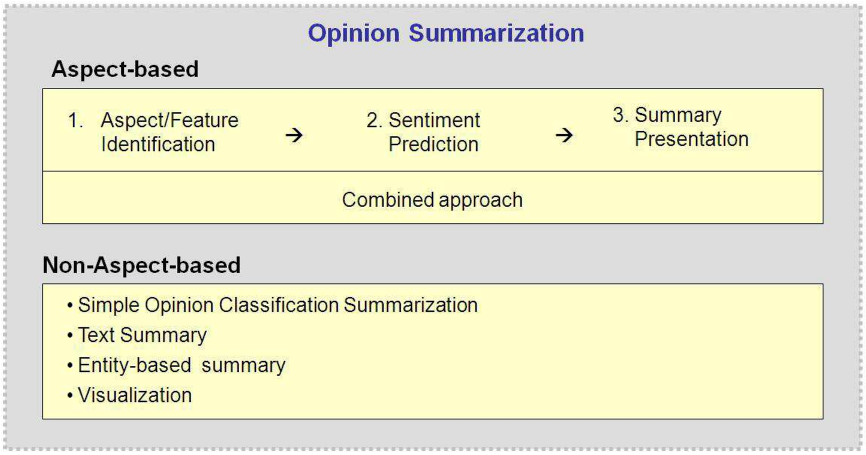


Fig. 2. Overview of opinion summarization techniques

very popular and have been heavily explored over the last few years [Hu and Liu 2004a; 2004b; 2006; Ku et al. 2006; Liu et al. 2005; Lu et al. 2009; Mei et al. 2007; Popescu and Etzioni 2005; Titov and McDonald 2008; Zhuang et al. 2006].

Non-aspect-based summarization includes all other kinds of opinion summarization works which do not divide the input texts into sub topics. The non-aspect-oriented summaries either assume that the opinion text has been pre-segmented by aspects or simply produce a generalized summary without consideration of aspects. Such approaches touch diverse concepts from text summarization to information visualization [Lu and Zhai 2008; Kim and Zhai 2009; Ganesan et al. 2010; Balahur and Montoyo 2008; Chen et al. 2006; Mishne et al. 2006; Stoyanov and Cardie 2006b; 2006a; 2008]. In the next sections, we will introduce the relevant approaches in each of these categories.

4. ASPECT-BASED OPINION SUMMARIZATION

The most common type of opinion summarization technique is *Aspect-based Opinion Summarization*. Aspect-based summarization involves generating opinion summaries around a set of aspects or topics (also known as features). These aspects are usually arbitrary topics that are considered important in the text being summarized. In general, aspect-based summarization is made up of three distinct steps - *aspect/feature identification*, *sentiment prediction*, and *summary generation*. Some approaches, however, integrate some of the three steps into a single model. Figure 4 shows brief explanation of the three steps in aspect-based summarization. The feature identification step is used to find important topics in the text to be summarized. The sentiment prediction step is used to determine the sentiment orientation (positive or negative) on the aspects found in the first step. Finally, the summary generation step is used to present processed results from the previous two steps more effectively.

Various methods and techniques have been proposed to solve challenges in each of these steps. In the following three subsections, we will describe core techniques used in the aspect/feature identification step, the sentiment prediction step, integrated approaches, and the summary generation step. Table I shows the techniques used in each step. The studies

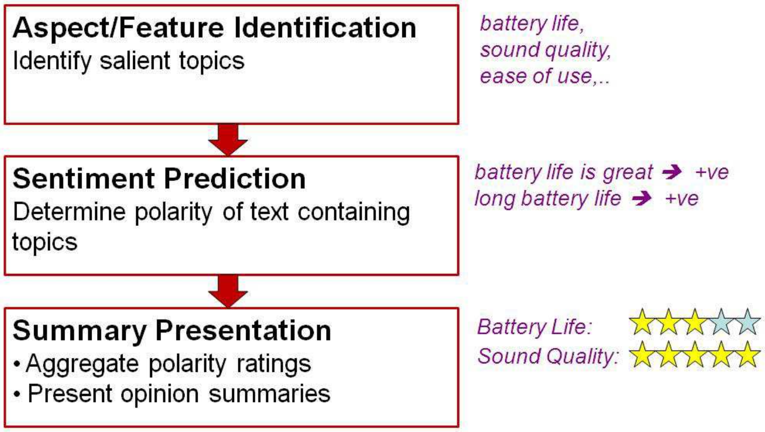

Aspect/Feature Identification
Identify salient topics
battery life,
sound quality,
ease of use,..
Sentiment Prediction
Determine polarity of text containing topics
battery life is great ➔ +ve
long battery life ➔ +ve
Summary Presentation
• Aggregate polarity ratings
• Present opinion summaries
Battery Life:
Sound Quality:

Fig. 3. General three steps of aspect-based opinion summarization

are ordered by the last name of the first author followed by the year of publication.

4.1 Aspect/Feature Identification.

Aspect/feature identification involves identifying salient topics within the text to be summarized. For example, if we want to generate an opinion summary about ‘iPod’, some of the common aspects are ‘battery life’, ‘sound quality’ and ‘ease of use’. The purpose of this step is to find these subtopics. In some cases, these topics are assumed to be known and hence this step is not required.

4.1.1 *NLP Techniques for Feature Discovery.* Most approaches [Lu et al. 2009; Popescu and Etzioni 2005; Hu and Liu 2004b; 2004a] attempt to identify features in the opinion text with the help of NLP-based techniques. Part-of-speech (POS) tagging and syntax tree parsing are very common starting points for feature discovery. For example, as aspects/features are usually noun phrases, even basic POS tagging allow people to find candidate aspects. The annotated opinion texts are then further analyzed using data/text mining techniques explained in Section 4.1.2.

In the recent work [Lu et al. 2009], shallow parsing was used to identify aspects for short comments. In short comments, most opinions are expressed in concise phrases, such as ‘well packaged’ and ‘excellent seller’. With this in mind, it is assumed that each phrase is parsed into a pair of head term and modifier, where the head term is about an aspect or feature, and the modifier expresses some opinion towards this aspect (e.g. ‘fast[modifier] shipping[head]’). The head terms in the text are then clustered to identify k most interesting aspects.

[Popescu and Etzioni 2005] used the KnowItAll system [Etzioni et al. 2004], a web-based domain-independent information extraction system, to extract explicit features for the given product class from parsed review data. This work used a more involved approach to extracting features compared to other works. First, the system recursively identifies both *parts* (e.g. scanner cover) and *properties* (e.g. scanner size) of the given product class until no more candidates are found. Then the system finds related concepts and extracts their parts and properties. To find parts and properties, noun phrases are extracted from

Table I. Techniques used in Aspect-based summarization

	Aspect/Feature Identification	Sentiment Prediction	Summary Generation
[Hu and Liu 2004a; 2004b; 2006]	**NLP-based Technique.** Perform POS tagging and generate n-grams. **Mining**. Use association rule mining to find all rules.	**Lexicon-based.** Use seed sentiment words and then use WordNet to generate more sentiment words.	**Statistical summary.** Sentiment distribution of each aspect with classified sentences. Graph representation proposed by [Hu and Liu 2006].
[Ku et al. 2006]	**Mining.** Use the frequency of terms in paragraphs and across paragraphs.	**Lexicon-based.** Use sentiment words to assign opinion scores to sentences.	**Text Selection.** Sentence selection based on TF-IDF scores of words **Summary with a timeline.** Show opinion changes over a timeline.
[Liu et al. 2005]	Same techniques used in [Hu and Liu 2004a; 2004b]	**Known in advance.** Orientation assigned to phrases based on whether it comes from 'Pros' or 'Cons' reviews.	**Statistical summary.** Opinion observer. Generate graph-based statistics with comparison of several products.
[Lu et al. 2009]	**NLP-based technique.** Identify head terms and cluster head terms into k interesting aspects.	**Learning-based technique**. Use overall ratings and a Naive Bayes classifier.	**Text Selection.** Show the most occurring phrase in each aspect. **Aggregated ratings.** Average sentiment rating of phrases within each aspect. **Text selection.** Choose phrases with highest support in each aspect.
[Mei et al. 2007]	**Integrated Approach.** Joint topic and sentiment modeling using Topic Sentiment Mixture (TSM). Model and extract multiple subtopics and sentiments in a collection of blog articles.		**Text selection.** Top scored sentence by topic modeling results. **Summary with a timeline.** Show opinion changes over a timeline.
[Popescu and Etzioni 2005]	**NLP-based technique.** Use KnowItAll system to extract features	**Other.** NLP-based technique: Dependency parsing to find heads and modifiers to discover opinion phrases. Statistical: Use relaxation labeling to predict sentiment orientation of opinion phrases.	**Text Selection.** Show the strongest opinion word for each aspect.
[Titov and McDonald 2008]	**Integrated Approach.** Joint topic and sentiment modeling using Multi-Grain LDA (MG-LDA). Extract ratable aspects using local and global topics.		**Text selection.** Top probability words for each topic.
[Zhuang et al. 2006]	**Other.** Use lexicon and regular expressions.	**Lexicon-based.** Use seed sentiment words and then use WordNet to generate more sentiment words.	**Statistical summary.** Sentiment distribution of each aspect class and corresponding sentences for each aspect and sentiment.

reviews, and the phrases that satisfy a minimum support are retained. Then the KnowItAll's Feature Assessor evaluates each noun phrase by computing PMI scores between the phrase and meronymy discriminators associated with the product class (e.g. of scanner, scanner has, scanner comes with, etc. for the Scanner class). Parts are then distinguished from properties using WordNet.

Shallow NLP approaches like POS tagging and parsing are quite effective for feature extraction as these techniques are well studied, and many state-of-the-art parsers and taggers are known to have high accuracies. One potential problem is the practicality of these approaches. The speed of parsing or tagging is still not 'efficient' enough for large scale processing. Also, such shallow NLP-based techniques may not be sufficient in discovering all the features. This is because features are not always nouns, and often times they are not explicitly specified in the text. For example, the sentence, 'The mp3 player is small', implicitly mentions the 'size' feature, but there is no mention of the word 'size' in the sentence. This may require some domain knowledge or help from some ontological word dictionary.

4.1.2 *Mining Techniques for Feature Discovery.* Another commonly used methods to identify features is a 'mining' approach [Archak et al. 2007; Popescu and Etzioni 2005; Hu and Liu 2004b; 2004a]. Frequent itemset mining can compensate the weaknesses of pure NLP-based techniques. This approach does not restrict that only certain types of words or phrases can become candidate features. Instead, other information like the *support* information is used to determine if a particular word or phrase is a feature or not. Certain non-promising features are even pruned with the use of mutual information and redundancy rules. This approach to feature discovery shows reasonable performance especially with product reviews.

[Hu and Liu 2004b; 2004a] used supervised association rule mining-based approach to perform the task of feature extraction. Their methods are based on the idea that each sentence segment contains at most one independent feature. First, each review sentence is divided into a set of sentence segments based on separation by '.', ',' , 'and', 'but', etc; then all the feature words are manually tagged. With the segmented and tagged data set, Association Rule Mining is performed to learn rules of the form $A_1 A_2 ... A_n \Rightarrow [feature]$ for predicting feature words, based on the remaining words in a sentence segment and their POS tags. Since association rule mining does not account for the order of $A_1, A_2 ... A_n$ in a sentence, many of the learnt rules can be pruned based on inconsistency of the patterns with English grammar. Features on a new input dataset are then extracted using these trained rules. In case two rules resulted in two different features for the same sentence segment, the more frequently occurring feature is chosen.

[Zhuang et al. 2006] used a slightly different approach for extracting features in movie reviews. Since many of the features in their case are around the cast of a movie, they build a feature list by combining the full cast of each movie to be reviewed. A set of regular expressions is then used to identify whether a word in a review matched one of the words in the feature list.

[Ku et al. 2006] introduced a fairly simple approach to discover features. They consider paragraph level frequencies as well as document level ones to help identify features. While most previous works used document frequency or just term frequencies within a document, this work analyzes frequencies across paragraphs and frequencies within paragraphs.

The study by [Archak et al. 2007] is different from other approaches so far. Their methods use a combination of text mining and econometric techniques. The methods attempt to first decompose product reviews into segments that evaluate the individual characteristics of a product (e.g., image quality and battery life for a digital camera). Then they adopt methods from the econometrics literature, specifically the hedonic [Rosen 1974] regression concept, to estimate: (a) the weight that customers place on each individual product feature, (b) the implicit evaluation score that customers assign to each feature, and (c) how these evaluations affect the revenue for a given product. By using product demand as an objective function, they derive a context-aware interpretation of opinions. Based on the analysis, they show how customers interpret the posted comments and how the comments affect customers' choices. The intuition here is that the results can be used by manufacturers to determine which features contribute most to the demand for their product. Such information can also help manufacturers facilitate changes in product design over the course of a product's life cycle.

One problem with mining-based approaches is that it may work differently in different domains. Sometimes, heuristics used for finding features need to be redefined for different domains. Also, parameters like the support threshold need to be tuned for different applications since a stable and uniform performance cannot be guaranteed with a global setting.

4.2 Sentiment Prediction

The feature discovery step is often followed by sentiment prediction on the text containing features that are previously found. Sentiment prediction in itself is an active research area. While there are many techniques solely for this task, in this section, we will discuss the techniques used within the framework of opinion summarization.

The purpose of sentiment prediction in the current context is to allow for the discovery of sentiment orientation (positive or negative) on the aspect/feature. Different people may have different views about similar aspects. For example, some people may find that the iPod's battery life 'is excellent', while others may find that it 'does not last long'. Thus, the results of aspect-based sentiment predictions would help users digest the general sentiments on the aspect.

4.2.1 *Learning-based Methods for Sentiment Prediction.*

Learning-based prediction can incorporate many features and formulate a problem as sentiment classification. All types of information can be potentially cast as features. Lexicons and rules which will be mentioned in Section 4.2.2 are one of the important features for learning-based predictions. By using characteristics of words around the target text, machine learning method even can capture context to some extent.

[Lu et al. 2009] is one of the few studies useing a learning-based strategy in aspect-based summarization. They propose two methods for classifying each phrase clustered into the k interesting aspects (see Section 4.1.1) into a rating $r(f)$. First they assume that the rating of each aspect is consistent with its overall ratings. In other words, each phrase mentioned in a comment shares the same rating as the overall rating of comments. With this assumption, the aspect ratings can be calculated by aggregating ratings of all the phrases about each aspect.

In the second method, instead of blindly assigning the same rate to each phrase as the overall rating of the comment, they learn aspect level rating classifiers using the global information of the overall ratings of all comments. Then each phrase is classified by the globally trained rating classifier. They essentially classify each phrase by choosing the rating class that has the highest probability of generating the modifier in the phrase, which is basically a Naive Bayes classifier with uniform prior on each rating class. The ratings are then aggregated by averaging the rating of each phrase within an aspect. This method of prediction is shown to work much better than just using the overall ratings.

While many studies on sentiment prediction use machine learning-based approaches, it is the least common approach within the context of opinion summarization. This was probably due to the difficulty in obtaining labeled examples to train a high accuracy classifier. Preparing big enough annotated data is a challenge in using learning-based methods. It is even harder to find a data for general domain, and the trained model in one domain may not work well in other domains.

4.2.2 *Lexicon/Rule-based Methods for Sentiment Prediction.* Lexicon-based sentiment prediction is very popular in the context of opinion summarization [Hu and Liu 2004b; 2004a; Zhuang et al. 2006; Ku et al. 2006]. This technique generally relies on a sentiment word dictionary. The lexicon typically contains a list of positive and negative words that are used to match words in the opinion text. For example, if an opinion sentence has many words from the positive dictionary, we can classify it as having a positive orientation. These word lists are often used in conjunction with a set of rules or can be combined with the results of POS tagging or parsing.

For identifying the opinions about features and their orientation, [Hu and Liu 2004b; 2004a] proposed a simple yet effective method based on WordNet. They start with a set of about 30 seed adjectives for each predefined orientation (positive and negative). Then they use the similarity and antonymy relations defined in WordNet for assigning positive or negative orientations to a large set of adjectives. Thus, the orientation of an opinion about a feature was decided by the orientation of the adjective around it.

Similarly, [Ku et al. 2006] used a set of positive and negative words to predict sentiments. They used two sets of sentiment words GI [2] and CNSD [3]. They enlarged the seed vocabulary using two thesauri Cilin [J. et al. 1982] and BOW [4]. The orientation of an opinionated sentence is decided based on the orientations of its words. Instead of using a set of rules, they assigned sentiment scores to sentences assigned to topics. These scores represent the sentiment degree and polarity. In addition to having a polarity of positive and negative, if certain words like 'say', 'present', and 'show' were present in the sentence, a zero opinion score was assigned as a neutral opinion.

[Zhuang et al. 2006] used dependency relationships to identify opinions associated with feature words. In order to identify the orientation of the opinions, they used a strategy similar to that of [Hu and Liu 2004b; 2004a]. They identified the top 100 positive and negative opinionated words from a labeled training set and then used WordNet synsets to assign orientations to other words. Furthermore, the orientation of a word was reversed if there was a negation relation such as 'not' or 'anti' involved.

This line of work is popular because it is faily simple and lexicons can be good features for learning-based methods. Lexicon-based approaches are known to work well in domains like product reviews where people are explicit about their expressions (e.g. 'The battery life

[2] http://www.wjh.harvard.edu/ inquirer/

[3] http://134.208.10.186/WBB/EMOTION_KEYWORD/ Atx_emtwordP.htm

[4] http://bow.sinica.edu.tw/

is bad'). However, in harder domains like movie reviews where people are often sarcastic, such a method yields in poorer performance because the context was often ignored. Also, the performance of this method depends on the quality of the dictionary used. For the best performance, different dictionaries have to be defined for different domains and aspects.

4.2.3 *Other Methods for Sentiment Prediction.* [Popescu and Etzioni 2005] used words in the vicinity of the features found as a starting point in predicting the sentiment orientation. Basic intuition is that an opinion phrase associated with a product feature tends to occur in its vicinity. Instead of using simple word window to check the words in vicinity, they use syntactic dependencies computed by MINIPAR [Lin 1998]. Heads and their corresponding modifiers in dependency parsing results are considered as potential opinion phrases.

They then use a well-known computer vision technique, relaxation labeling [Hummel and Zucker 1987], to predict the polarity of extracted opinion phrases. Relaxation labeling uses an update equation to re-estimate the probability of a word label based on its previous probability estimate and the features of its neighborhood. The initial probability is computed using a version of Turney's PMI-based approach [Turney and Littman 2003]. This technique is found to generate opinions and its corresponding polarity with high precision and recall. However, this is tested only on user reviews in the products domain, so it may not be general enough to be used in any arbitrary domain. In addition, since the sentiment prediction step alone is multi-faceted and very involved, the approach can have scalability issues.

4.3 Integrated Approaches

Some studies in aspect based summarization do not have a clear separation of the summarization steps explained earlier. We refer to these approaches as integrated approaches [Mei et al. 2007; Titov and McDonald 2008] which mainly use probabilistic mixture models namely PLSA [Hofmann 1999] and LDA [Blei et al. 2003].

4.3.1 *Topic Sentiment Mixture Model.* Before we introduce Topic Sentiment Mixture model, it is necessary to describe a general modeling approach-based topic finding task. Now, let us assume that we have reviews on similar products and these reviews come from multiple sources. We would like to know what the common themes are across the different sources and also want to get an idea of what each specific source talks about. For such a task, [Zhai et al. 2004] proposed a nice way of doing this with an approach called *Comparative Text Mining* (CTM). The task in this approach was to discover any latent common themes across all collections as well as summarize the similarity and differences of these collections along each common theme. This is done by simultaneously performing cross-collection clustering (to obtain the common theme) and within-collection clustering (to obtain the differences). This approach focuses on finding coherent summaries on different topics. Also, this approach is general enough that we can use CTM on any type of text collection. However, this approach is not for opinionated collections, so it does not have consideration in sentiment analysis.

For opinion summarization, by adding sentiment models to CTM, [Mei et al. 2007] processed sentiment prediction and aspect identification in one step with topic modeling. In this approach, they use PLSA to capture the mixture of topics and sentiments simultaneously. The propose *Topic Sentiment Mixture* (TSM) model (Figure 4.3.1) which can reveal the latent topical facets in the collection of text and their associated sentiments. TSM

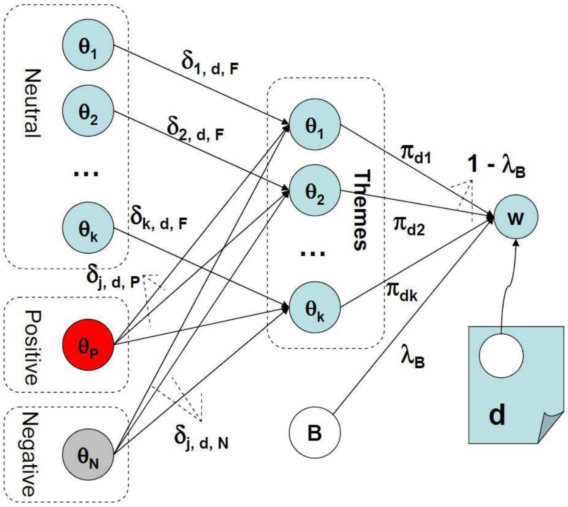


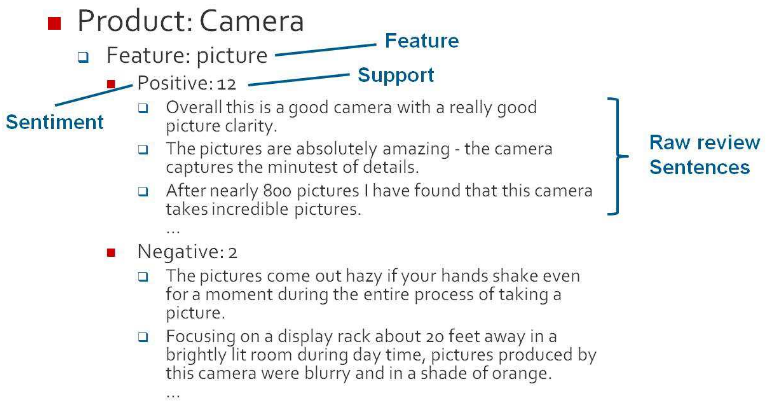
Fig. 4. Generation process of topic sentiment mixture model [Mei et al. 2007]. d: document, w: word, B: background model, θ: topic model, π_{dj}: the probability of choosing jth topic in document d, $\delta_{j,d,F/P/N}$: sentiment coverage of topic j in document d.

has an additional level on top of CTM. They assume that topics are generated by one of the neutral topics and one of positive and negative topics. By setting basic positive and negative sentiment words as priors, they let positive and negative topics work as intended.

Apart from obtaining summaries based on topics and sentiments, they also design a Hidden Markov Model (HMM) structure to utilize the sentiment models and topic models estimated with TSM to extract topic life cycles and sentiment dynamics. The TSM model is quite general in that it can be applied to any text collections with a mixture of topics and sentiments. Thus, it has many potential applications, such as search result summarization, opinion summarization, and user behavior prediction.

4.3.2 *Multi-Grain LDA Model.* A recent work by Titov and McDonald [Titov and McDonald 2008] jointly modeled text and aspect ratings. In this work, the key idea is to find ratable aspects within texts on a given topic, and then find representative summaries for each aspect. For example, once an aspect such as *staff* is found, representative summaries like *waitress* and *bartender* become part of the *staff* topic. The approach uses a model based on extensions to the standard topic modeling method, LDA, to induce multi-grain topics. This approach proposes a multi-grain model as being more appropriate for this sort of a task since standard models tend to produce topics that correspond to global properties of objects (e.g., the brand of a product type) rather than the aspects of an object that tend to be rated by a user. The proposed *Multi-Grain LDA* (MG-LDA) models two distinct types of topics: global topics and local topics. The hypothesis is that ratable aspects are captured by local topics, and global topics capture properties of reviewed items. The main goal of this work is to use rating information to identify more coherent aspects.

MG-LDA does not directly predict sentiment orientation. However, because they generate features using rating information, it can be potentially applied to other models that consider sentiments such as TSM. Therefore, we categorize this study into the integrated



- Product: Camera
  - Feature: picture
    - Positive: 12
      - Overall this is a good camera with a really good picture clarity.
      - The pictures are absolutely amazing - the camera captures the minutest of details.
      - After nearly 800 pictures I have found that this camera takes incredible pictures.
      - ...
    - Negative: 2
      - The pictures come out hazy if your hands shake even for a moment during the entire process of taking a picture.
      - Focusing on a display rack about 20 feet away in a brightly lit room during day time, pictures produced by this camera were blurry and in a shade of orange.
      - ...

Fig. 5. Feature-based textual statistical summary. Raw review sentences are classified by sub features and sentiment orientations.

approach section.

4.4 Summary Generation

Using the results of feature discovery and sentiment prediction, it is then critical to generate and present the final opinion summaries in an effective and easy to understand format. This typically involves aggregating the results of the first two steps and generating a concise summary.

In the following subsections, we will describe various generation methods for opinion summarization. While each technique has its own focus, some techniques can be combined with others. For example, we may add a timeline to text selection methods.

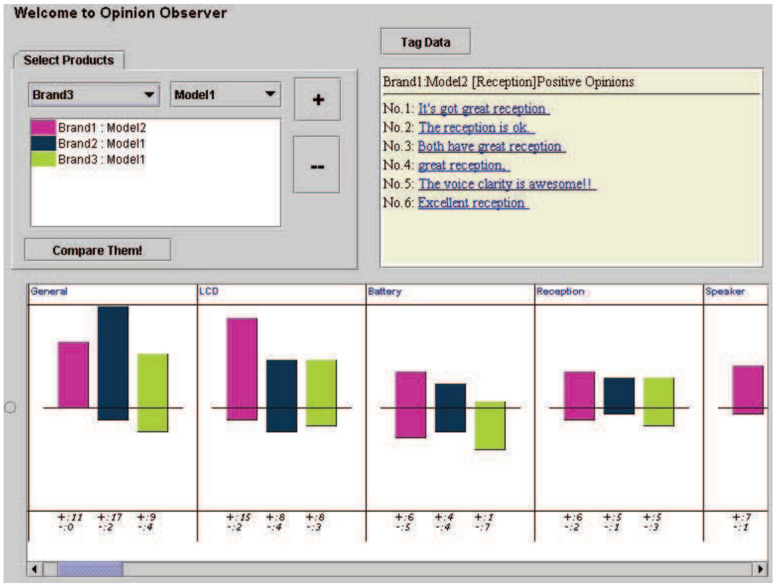
4.4.1 *Statistical Summary.* While there are various formats of summaries, the most commonly adopted format is a summary showing statistics introduced by [Hu and Liu 2004b; 2004a; 2006; Zhuang et al. 2006]. Statistical summary directly uses the processed results from the previous two steps - a list of aspects and results of sentiment prediction. By showing the number of positive and negative opinions for each aspect, readers can easily understand the general sentiments of users at large. Along with the positive and negative occurrences, all sentences with sentiment prediction in each aspect is shown (Figure 4.4.1).

[Hu and Liu 2006] showed statistics in a graph format. With the graph representation, we can obtain people's overall opinions about the target more intuitively. [Liu et al. 2005] developed software, Opinion observer, which shows statistics of opinion orientation in each aspect and even enables users to compare opinion statistics of several products. An example result is shown in Figure 4.4.1, which compares opinions on three cell phones from three different brands.

This format of summary has been widely adopted even in the commercial world. Figure 7 shows a sample structured summary used on Bing [5].

4.4.2 *Text Selection.* While statistical summaries help users understand the overall idea of people's opinion, sometimes reading actual text is necessary to understand specifics.

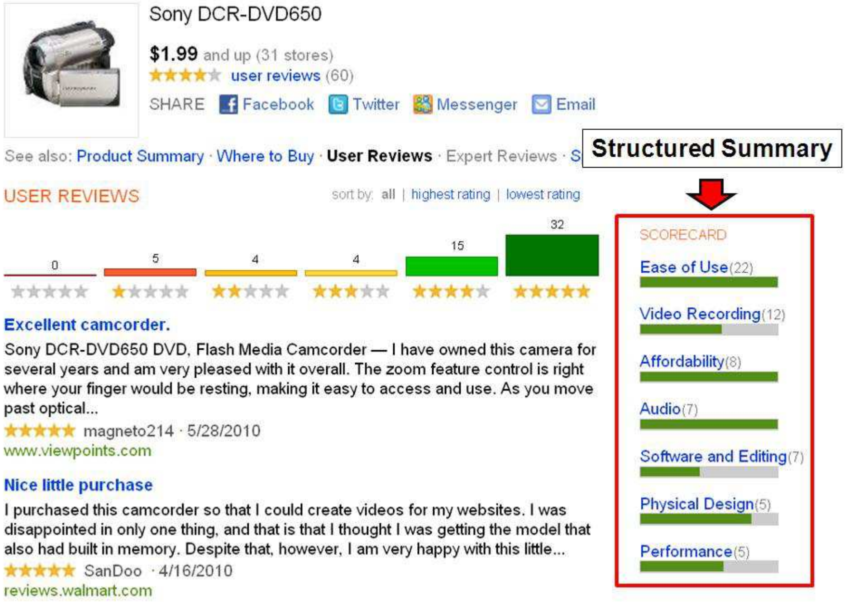
[5] http://www.bing.com



Brand1:Model2 [Reception]Positive Opinions

No.1: It's got great reception
No.2: The reception is ok.
No.3: Both have great reception
No.4: great reception.
No.5: The voice clarity is awesome!!
No.6: Excellent reception

Fig. 6. Visualization of an aspect summary by [Liu et al. 2005]



Excellent camcorder.

Sony DCR-DVD650 DVD, Flash Media Camcorder — I have owned this camera for several years and am very pleased with it overall. The zoom feature control is right where your finger would be resting, making it easy to access and use. As you move past optical...

★★★★★ magneto214 · 5/28/2010
www.viewpoints.com

Nice little purchase

I purchased this camcorder so that I could create videos for my websites. I was disappointed in only one thing, and that is that I thought I was getting the model that also had built in memory. Despite that, however, I am very happy with this little...

★★★★★ SanDoo · 4/16/2010
reviews.walmart.com

Fig. 7. Structured summary on Bing Product Search [6]

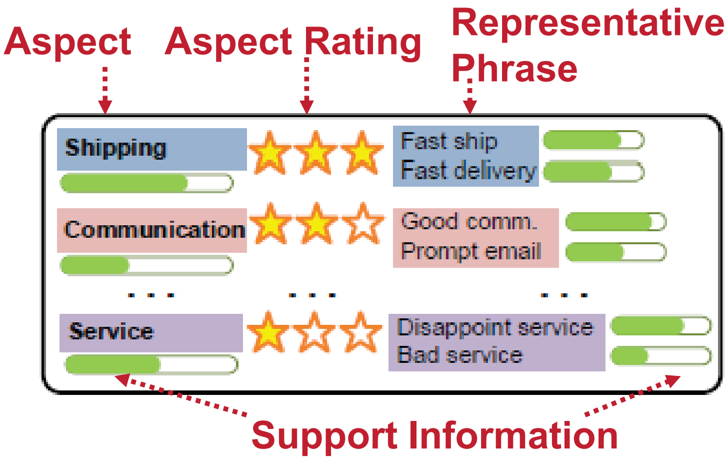


Fig. 8. Structured summary generation by Lu et al. [Lu et al. 2009]

Due to the large volume of opinions on one topic, showing a complete list of sentences is not very useful. To solve this problem, many of the recent studies [Titov and McDonald 2008; Popescu and Etzioni 2005; Lu et al. 2009; Mei et al. 2007; Ku et al. 2006] try to show smaller pieces of text as the summary. They use different granularities of summaries including word, phrase and sentences level granularities.

With the topic modeling methods, a word level summary are usually provided for each topic [Titov and McDonald 2008] because the list of words and their probability is a natural output of the topic modeling approaches. [Popescu and Etzioni 2005] also used word selection as the summary. They rank opinion words associated to features and show the strongest opinionated word for each aspect. Going beyond, word level summaries, [Lu et al. 2009] show that is is possible to generate short representative phrases (that are high occurring) using clustering approaches. This approach however is only tested on eBay comments, which are rather short to start with.

A sentence level summary can provide a deeper level of understanding of a topic. [Mei et al. 2007] score the probability of each sentence to each topic using word probability in topic modeling of TSM model. By choosing the top ranked sentence in each category, they are able to show the most representative sentence. [Ku et al. 2006] on the other hand, score sentences based on the TF-IDF of their words and select the most relevant and discriminative sentence to be shown as summary.

4.4.3 *Aggregated Ratings.* [Lu et al. 2009] proposed the advanced summary, *aggregated ratings*, which combines statistical summary and text selection. Based on the discovered aspects using clustering and topic modeling, they average the sentiment prediction results of phrases for each aspect as the final sentiment rating for that aspect. Aspect ratings are shown with representative phrases. Figure 8 shows an example of their summary generation.

4.4.4 *Summary with a Timeline.* [Ku et al. 2006; Mei et al. 2007] showed opinion trends over a timeline. General opinion summarization focuses on finding statistics of the 'current' data. In reality, opinions change as time goes by. Opinion summary with a timeline helps us see the trend of opinions about a target easily, and it also can tell us ideas for further analysis. To figure out what changes people's opinions, we can analyze the events that happened at the drastic opinion change. For example, Figure 4.4.4 shows the change of opinions towards four election candidate, and we can easily identify that there is a drastic opinion change on the Election Day.

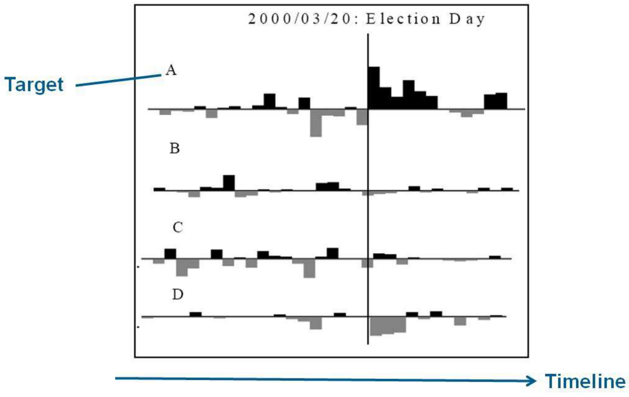


Fig. 9. Opinion summary with a timeline. Show the change of opinions towards four targets along the timeline [Ku et al. 2006].

5. NON-ASPECT-BASED OPINION SUMMARIZATION

In addition to studies we introduced in previous sections, there are many opinion summarization works which do not fit into the aspect-based summary format. They are not bound by the aspect-based format and suggest different formats for opinion summarization. Some of them can be combined together or incorporated into aspect-based summarization methods. We have categorized them into basic sentiment summarization, advanced text summarization, visualization, and entity-based summarization. Following is the summary of non-aspect-based summarization methods.

(1) Basic Sentiment Summarization

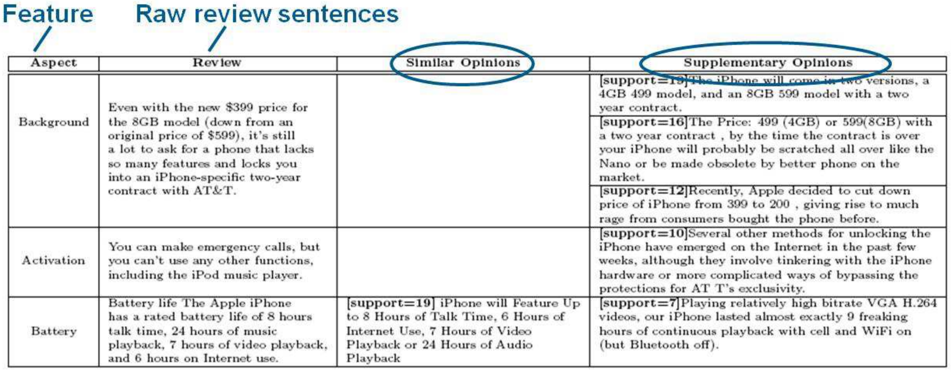
(2) Text Summarization
  (a) Opinion Integration [Lu and Zhai 2008]
  (b) Contrastive Opinion Summarization [Kim and Zhai 2009]
  (c) Abstractive Text Summarization [Ganesan et al. 2010; Ganesan et al. 2012]
  (d) Multi-lingual Opinion Summarization [Balahur and Montoyo 2008]

(3) Visualization [Chen et al. 2006; Mishne et al. 2006]

(4) Entity-based summary [Stoyanov and Cardie 2006b; 2006a; 2008]

5.1 Basic Sentiment Summarization

Using the prediction results from sentiment classification (Section 2.1), basic sentiment summary can be generated. Sentiment classification decides the sentiment orientation of the input texts per classification unit (sentence, document, etc.). By simply counting and


Feature
Raw review sentences

Aspect	Review	Similar Opinions	Supplementary Opinions
Background	Even with the new $399 price for the 8GB model (down from an original price of $599), it's still a lot to ask for a phone that lacks so many features and locks you into an iPhone-specific two-year contract with AT&T.		[support=19]The iPhone will come in two versions, a 4GB 499 model, and an 8GB 599 model with a two year contract.
			[support=16]The Price: 499 (4GB) or 599(8GB) with a two year contract , by the time the contract is over your iPhone will probably be scratched all over like the Nano or be made obsolete by better phone on the market.
			[support=12]Recently, Apple decided to cut down price of iPhone from 399 to 200 , giving rise to much rage from consumers bought the phone before.
Activation	You can make emergency calls, but you can't use any other functions, including the iPod music player.		[support=10]Several other methods for unlocking the iPhone have emerged on the Internet in the past few weeks, although they involve tinkering with the iPhone hardware or more complicated ways of bypassing the protections for AT T's exclusivity.
Battery	Battery life The Apple iPhone has a rated battery life of 8 hours talk time, 24 hours of music playback, 7 hours of video playback, and 6 hours on Internet use.	[support=19] iPhone will Feature Up to 8 Hours of Talk Time, 6 Hours of Internet Use, 7 Hours of Video Playback or 24 Hours of Audio Playback	[support=7]Playing relatively high bitrate VGA H.264 videos, our iPhone lasted almost exactly 9 freaking hours of continuous playback with cell and WiFi on (but Bluetooth off).

Fig. 10. Opinion integration example [Lu and Zhai 2008]. Provide similar and supplementary opinions from non-expert reviews (right two columns) to the structured opinions from expert reviews (left two columns).

reporting the number of positive opinions and negative opinions, we can easily generate a simple statistical opinion summary.

This summary can show the overall opinion distribution of input data set without sophisticated aspect identification steps. However, this type of summarization only can show sentiment analysis results at a very coarse granularity. While the format used in simple opinion classification has been widely adopted, such a summary may not be sufficient enough to help people understand the specifics within the opinions. This motivates studies on aspect-based summarization and textual summaries.

5.2 Text Summarization

5.2.1 *Opinion Integration.* [Lu and Zhai 2008] used different strategies to process texts depending on the type of sources. (Figure 5.2.1). They divide opinion documents into two categories, expert opinions and ordinary opinions. Expert opinions are articles which is usually well structured and easy to find features. For example, CNET expert reviews or Wikipedia articles are expert opinion articles. Although expert opinions are pretty complete by itself, they are not updated often; therefore, they do not usually reflect latest changes immediately. Ordinary opinions are the other unstructured articles. Most of the private blog articles and user reviews are considered ordinary opinions. They may have unimportant information, but they tend to be updated more often; therefore, they reflect recent news very well.

Opinion integration is for integrating these two kinds of sources and getting a complete opinion summary. First, they extract structured information (aspect and feature data) from the expert opinions to cluster general documents. By using a semi-supervised topic model using PLSA technique, they take advantage of two different sources. Features extracted from the expert opinions are used as prior to the second step that is to analyze ordinary opinions. Similar opinions were integrated into the expert reviews, and information which was not covered by the expert opinions for each aspect was added into the summary as supplementary opinions. In addition, even information from ordinary opinions about aspects which is not covered by the expert opinions are added to the summary as supplementary opinion on extra aspects. Because we can insert any kinds of expert opinions as input, this task can be applied to any domain.

The proposed approach uses one type of expert review as an input. However, there can be conflicts among different expert reviews with different aspect structures. Data cleaning and alignment method would be helpful to analyze and combine structures of various expert reviews; then, we would be able to have more complete list of aspects.

5.2.2 *Contrastive Opinion Summarization.* Existing opinion summaries usually generate two sets of sentences; positive and negative. Aspect-based methods further divide sentences according to subfeatures. However, users still need to read divided sentences to understand the details opinions. Especially, there can be sentences with mixed orientation which are difficult to be classified clearly. For example, two sentences, 'The battery life is long when we rarely use buttons' and 'The battery life is short when we use buttons a lot' would be classified as positive and negative respectively, but they are actually saying the same fact.

[Kim and Zhai 2009] proposed a method to show contrastive opinions effectively, *Contrastive Opinion Summary* (COS). COS further summarizes the output of the existing opinion summary. Given positive and negative sentences as inputs, COS generates contrastive paired sentences. To be a good contrastive summary, generated sentence pairs should be representative of input sentences as well as contrastive to show contradiction more effectively. They formulate the problem as an optimization framework and propose two approximation methods for generating representative and contrastive sentence pairs. Representative-first approximation method clusters each positive and negative sentence set into k clusters and finds contrastive pairs. Contrastive-first approximation method finds contrastive pairs first and selects representative pairs among them. They mainly use similarity functions based on word overlap and also experimented with the variation like semantic similarities between words for similarity functions.

This work suggest new summarization problem, contrastive opinion summarization. By further summarizing already classified sentences, it decreases the volume of data that users should read. In addition, by showing contrastive pairs, it catches the important points we should compare more effectively. However, the basic techniques used in COS are simple. They mainly use the word overlap-based content similarity functions. By using more sophisticated NLP techniques, the accuracy of the algorithm can be improved. For example, tree alignment distance measure can be used for sentence similarity techniques. Also, they can select sentiment words more carefully than just choosing adjectives and negation words for contrastive similarity measure.

5.2.3 *Abstractive Text Summarization.* As explained in Section 2.3, because of the difficulties in text regeneration, abstractive summary is a less common strategy in text summarization than extractive summaries. Most of the techniques in opinion summarization specifically use simple keyword/phrase extraction or extractive sentence selection methods.

As opinions are highly redundant and scattered, extractive summarization methods may not capture all the major opinions if the wrong set of sentences are selected as summary sentences. This becomes especially crucial when there is a limit on the summary size where only a small number of sentences may be selected. In this case it would be hard to capture sentences that summarize all the opinions. Also, extractive methods tend to be quite verbose and this would make it not suitable for viewing on smaller screens. With this, Ganesan et al. proposed several *abstractive opinion summarization methods* [Ganesan et al. 2010; Ganesan et al. 2012] with the intuition that abstractive methods may be better suited at capturing the major opinions in text.

The first method proposed, called Opinosis [Ganesan et al. 2010], uses a novel graph-based summarization framework. In Opinosis, the first step is to generate a textual word graph (called the Opinosis-Graph) of the opinionated content, where each node represents a word and POS combination, and an edge represents the link between two words in a sentence. Using three unique properties of the graph data structure (redundancy capture, collapsible structures & gapped subsequences), various promising subpaths in the graph that act as candidate summaries are scored and short-listed. The top candidates are then used to generate the final Opinosis summaries.

The second approach [Ganesan et al. 2012] uses an optimization framework with an objective function capturing *representativeness* and *readability* along with constraints that ensure compactness of a generated summary. The starting point is essentially a set of high frequency unigrams from the text to be summarized. Each of these high frequency unigrams is then paired with every other unigram to form bigrams. For example, if we have the words battery and life, the bigrams generated would be battery life and life battery. Each bigram is then 'grown' into higher order n-grams and at each point the phrases generated are scored based on their representativeness and readability, where the phrases that do not satisfy the minimum scores would be discarded and prevented from growing further. The nal step is to sort all the candidate n-grams based on their objective function values which is the sum of the readability and representativeness scores.

It was shown that with both approaches it is possible to generate readable and concise textual opinion summaries without redundant information. Since both approaches are domain-independent, syntax-lean, and no training corpora needed, these approaches are fairly practical and general. However, due to the reliance on the surface order of words in the text, the semantic similarity between sentences is not captured. For example, a sentence like 'very good battery life' and 'fantastic battery life' would be considered two separate sentences with different meanings. To further improve on this, a deeper level of natural language understanding would be required.

5.2.4 *Multi-lingual Opinion Summarization.* As the different aspect, *Multi-lingual Opinion Summarization* [Balahur and Montoyo 2008] tried to introduce opinion summarization in to translation. Pre-processing steps are similar to general opinion summarization techniques. After analyzing features from English texts, they map concepts to Spanish using EuroWordNet.

This technique completely relies on EuroWordNet; therefore, the performance of the system is also completely depended on the performance of WordNet. Because WordNet does not include all the words and may even have errors, connecting words by utilizing general web information can be a possible solution.

5.3 Visualization

While most of previous introduced works showed summarization as a simple table-based statistical summary with representative text snippets or sentences, there were also attempts to show summary result in different ways to give more intuition to readers and increase readability.

[Chen et al. 2006] showed opinions on one topic with different graph structures. They

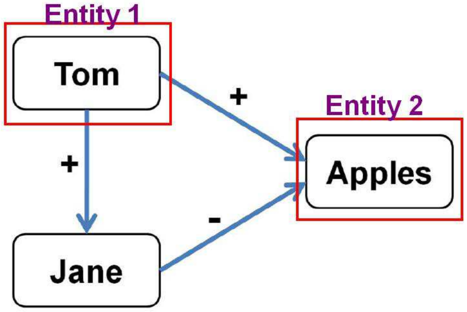


Fig. 11. Example entity-based summary

present term clusters with polarity information, words coordination, and decision tree-based review representation.

MoodViews [Mishne et al. 2006] is a visualization tool for blog mood analysis. There are three sub applications, Moodgrapher, Moodteller, and Moodsignals. Moographer is for showing aggregated mood level based on mood tags by users, and Moodteller is a similar tool using natural language processing steps for mood finding. MoodSignal is for finding reasons of special events which are represented as spikes in the mood aggregation graph.

A variety of analysis aspects can help to understand opinion distribution and characteristics. For example, users easily can tell whether there are more positive opinions than negative ones or not. Visualization is useful not only for public users who just want to understand opinions, but also for researchers who need to obtain better intuition to the summarization results. For example, from Chen et al.'s work [Chen et al. 2006], a polarity term distribution graph shows large term variation in negative opinions. By analyzing the reason of the phenomenon, researchers could find that there are more specific explanations for criticism in negative opinions; as a result, classification performance can be different in positive and negative opinions.

5.4 Entity-based summary

[Stoyanov and Cardie 2006b; 2006a; 2008] introduced a different type of summary, *Entity-based Summary*, which focused on 'who' talks what to 'whom'. The entity-based summary shows the entities in text and their relationships with opinion polarity annotations. This type of summary is also called 'term-based summary'. The summary is composed of opinion sources, targets, and opinions of sources to targets. For example, there are three sentences. "Tom likes apples. Jane hates apples. Tom loves Jane." For the first sentence, 'Tom' is a source, 'apples' is a target, and 'like' shows the opinions of 'Tom' to 'apples'. This relation also can be represented as the diagram; source and target entities are connected by an arrow annotated by opinion polarity. For the graphical summary of example sentences will be like the Figure 5.4.

For entity-based summary, finding and managing entities is a key issue because entities are used as main sources and targets. In actual texts, because many entities are referred by references such as a pronoun, the problem to find the correct referent, that is, coreference resolution, has been popularly studied as the first step.

[Stoyanov and Cardie 2006b; 2006a] proposed coreference resolution techniques to link sources mentioning the same entity. They use both mined rules for opinion data set and general noun phrase coreference resolution techniques. [Stoyanov and Cardie 2008] proposed a method for topic/target identification problem using coreference resolution.

The entity summarization is suggested but not all the subtasks are completely studied yet. As an initial step, only coreference resolution has been intensely studied. For the complete summary, we need other techniques for the next steps such as opinion identification, polarity determination, and ranking opinions.

Moreover, adding intensity of opinions can increase the expressivity of proposed opinion summary. Previous entity summarization shows only the binary polarity of opinion from one entity to the other. However, there can be different intensities in opinions. For example, we can think that 'love' has stronger positive opinion than 'like'. For expressing intensity in the graph summary, we may add more notations. For example, the thickness of arrow may express the intensity of opinions. Or using multiple number of + and - can show the intensity of opinion.

6. EVALUATION OF OPINION SUMMARIZATION

In this section, we will discuss various aspects of opinion summarization evaluation; how they obtain data set, which measures are used, and whether there are any public demo systems.

6.1 Data Set

The basic requirement of opinion summarization is the need for opinionated data. Popular data sets are reviews and blog articles on the web. Many researchers obtain data set by themselves by crawling target web sites with specific queries [Hu and Liu 2004a; 2004b; 2006; Lu et al. 2009; Kim and Zhai 2009; Titov and McDonald 2008; Zhuang et al. 2006]. Review sites are popularly crawled; for example, Amazon, eBay, and CNET for product reviews, TripAdvisor.com for hotel reviews, and IMDB for movie reviews are used.

Some researchers directly queried search engines, especially blog search engines, and crawled the results pages for evaluation [Ku et al. 2006; Mei et al. 2007; Kim and Zhai 2009; Titov and McDonald 2008; Lu and Zhai 2008].

There are some standard data sets that are commonly used to evaluate the task of opinion summarization; TREC [7], NCTIR [8], and MPQA [9]. These data sets are initially designed for sentiment classification. Because there is no standard data set specially designed for opinion summarization, not many researchers used standard data sets.

Making gold standard data to compare with system results is another issue in evaluation. Most of previous opinion summarization researches relies on human annotations [Hu and Liu 2004a; 2004b; 2006; Ku et al. 2006; Lu et al. 2009; Popescu and Etzioni 2005; Zhuang et al. 2006; Lu and Zhai 2008; Ganesan et al. 2010; Stoyanov and Cardie 2006b; 2006a; 2008]. Usually they provided data set and instructions to several human annotators to make labels. Most studies also show agreement rate among human assessors in evaluation section. Other than human annotations, [Lu et al. 2009; Titov and McDonald 2008] used ratings in the review data as a gold standard. [Mei et al. 2007] used output of other system,

[7] http://trec.nist.gov

[8] http://research.nii.ac.jp/ntcir/index-en.html

[9] http://nrrc.mitre.org/NRRC/publications.htm

Table II. List of publicized data set used for the previous research

	Data Set
[Hu and Liu 2004a; 2004b; 2006]	http://www.cs.uic.edu/ liub/FBS/sentiment-analysis.html
[Ku et al. 2006]	http://nlg18.csie.ntu.edu.tw:8080/opinion/index.html
[Kim and Zhai 2009]	http://sifaka.cs.uiuc.edu/ir/data/cos
[Ganesan et al. 2010]	http://sifaka.cs.uiuc.edu/ir/data/opinosis

Opinmind which is commercialized now [10], as a gold standard.

Some researchers make their data set and annotation used in evaluation public. Table II shows the list of publicized data set used for previous researches. [Hu and Liu 2004a; 2004b; 2006] posted product review data set with feature and polarity annotation on their web site [11]. [Ku et al. 2006] posted annotated opinion corpora [12]. Data set and annotation used for contrastive opinion summarization [Kim and Zhai 2009] [13] and abstractive opinion summarization [Ganesan et al. 2010] [14] are also publicized on their project pages.

6.2 Measures

The most popular evaluation measures are precision and recall. F-score are also used when authors want to show general performance combining precision and recall [Hu and Liu 2004a; 2004b; 2006; Ku et al. 2006; Lu et al. 2009; Popescu and Etzioni 2005; Zhuang et al. 2006; Kim and Zhai 2009; Stoyanov and Cardie 2006b; Balahur and Montoyo 2008]. In case of aspect-based opinion summarization, because the task of each step is fairly well divided, most studies evaluate each step separately. The main task of opinion summarization is finding appropriate information from input data. For example, feature identification is necessary in aspect-based summarization, and text selection method should choose the right contents in summary generation. Therefore, recall is used as one of the main measures. Measuring precision is also used in many subtasks in opinion summarization. First, any tasks using classification can be evaluated by precision. Especially, precision is a good method for evaluating sentiment classification which is one of the most important subtasks in opinion summarization. Also, precision is a good method to measure the performance when clustering or assigning texts into some categories/aspects.

Entity based opinion summarization methods [Stoyanov and Cardie 2006b; 2006a; 2008] additionally use evaluation measures for coreference resolution which is the main task. B-CUBED [Bagga and Baldwin 1998], CEAF [Luo 2005], and Krippendorff's α [Krippendorff 2003; Passonneau 2004] are representative coreference resolution evaluation measures.

In case proposed summarization methods generate some score and ranking of results, rank correlation and rank loss are used as evaluation measures [Lu et al. 2009; Titov and McDonald 2008]. Rank correlation is a measure comparing difference between two ranks. Kendal's τ and Spearman's rank correlation coefficient are representative measures. Rank loss [Snyder and Barzilay 2007] measures the average distance between the true and predicted numerical ratings. Because both [Lu et al. 2009] and [Titov and McDonald 2008]

[10] http://www.adaramedia.com

[11] http://www.cs.uic.edu/ liub/FBS/sentiment-analysis.html

[12] http://nlg18.csie.ntu.edu.tw:8080/opinion/index.html

[13] http://sifaka.cs.uiuc.edu/ir/data/cos

[14] http://sifaka.cs.uiuc.edu/ir/data/opinosis

used actual ratings of reviews obtained from the web as a gold standard, they need to compare actual values.

There are other quantitative measures. To evaluate how well generated summary represents the original inputs, coverage is used as a measure [Lu and Zhai 2008; Kim and Zhai 2009]. [Mei et al. 2007] use KL divergence to compare the generated model with the gold standard model. [Ganesan et al. 2010] use ROUGE [Lin 2004] which is a popular evaluation measure for general text summarization.

In addition to rigorous quantitative evaluation, qualitative observations are widely used to analyze example results [Mei et al. 2007; Lu and Zhai 2008; Kim and Zhai 2009; Titov and McDonald 2008]. The best evaluation would be human observation to all cases. However, because of limited resources, usually researchers design interesting usage scenarios and shows usefulness by analyzing one or two results in detail. Due to its own characteristics, usually, evaluation of visualization works are limited to showing screen shots of generated visual [Chen et al. 2006; Mishne et al. 2006]

Although opinion summarization is becoming active areas, as we can see above, there are not standardized data sets and evaluation measures for opinion summarization. Many of research studies use small data set crawled by themselves, and evaluation is performed in each sub task separately, not in the entire picture. We will discuss this problem more in Section 7.3.

6.3 Toolkits and Systems

To maximize the usefulness of research results, researchers tries to implement systems based on their studies. Some researches even publicize their demo system on the web.

[Liu et al. 2005] propose Opinion observer, which generates statistical opinion summary in each aspect. Opinion observer helps users to compare different products easily. Also, Opinion observer provides tagging interface for human annotators.

MoodViews [Mishne et al. 2006] [15] shows various demo applications based on blog opinion analysis. Basic application simply shows sentiment changes over a timeline (Moodgrapher), and the more sophisticated system even finds unusual peak in mood (Moodsignals).

[Kim and Zhai 2009] implemented two web demo systems [16]. First system is a simple implementation of comparative opinion summarization which users can try with their own inputs. The second system, Ooops!, is an opinion search engine with various opinion summarization features such as opinion filtering, graph-based opinion summary generation, opinion trend summary, and comparative opinion summary generation.

[Chen et al. 2006] show various possible visualization methods in their paper with screen shots. [Ganesan et al. 2010] posted executable demo system on their project page [17].

7. SUMMARY AND OPEN CHALLENGES

Because of its usefulness and needs from the people, opinion mining became an active research area. As the volume of the opinionated data increases, analyzing and summarizing opinionated data is becoming more important. To satisfy these needs, many kinds of opinion summarization techniques are proposed. Probabilistic approaches using statistics

[15] http://moodviews.com

[16] http://sifaka.cs.uiuc.edu/ir/data/cos

[17] http://sifaka.cs.uiuc.edu/ir/data/opinosis

of terms and heuristic approaches using predefined rules are representative works. In addition to the general opinion summarization studies, we introduced other types of opinion summarization works such as advanced text summarization, visualization, and entity-based summarization.

Opinion summarization started with heuristic methods. Many of heuristic rule-based methods showed a reasonable performance and let people know the usefulness of opinion summarization. In recent years, various probabilistic models such as PLSA and LDA started to be used. Most of integrated approaches were probabilistic model methods. Recent works such as opinion integration also used the probabilistic modeling approach.

Despite of a lot of research efforts, current opinion summarization studies still have many limitations and margins for improvement. In the previous subsections, we briefly mentioned limitations and possible future directions of some techniques. In this section, we will discuss the open challenges for overall opinion summarization area.

7.1 Improve Accuracy

Despite of a lot of research works, there are not many commercial services using opinion summarization. Although there are some services using opinion summarization, many of subtasks in the services are done by human effort.

One of the main reasons that companies cannot take advantage of opinion summarization techniques is their unsatisfactory accuracy. Unlike academic research works, commercial services require high accuracy. They cannot just provide users with erroneous services. However, because of difficulties in understanding language, opinion summarization does not support high accuracy enough to satisfy customers yet. There are a lot of complex sentences and vague expression to analyze. The accuracy problem is even worse in informal articles. Private blogs are one of the most important opinion sources. However, unlike expert articles, due to slangs, emoticons, and typos, personal articles have a lot of difficulties to analyze them automatically.

To solve this problem, techniques in each step should be improved. Because opinion analysis also starts from the understanding the meaning of texts, deeper NLP may help to improve accuracy. Otherwise, if it is difficult to develop a good general technique, specializing in one domain and using the domain knowledge can be one way to overcome the accuracy problem. As vertical search engine researches are becoming popular in the general information retrieval area, opinion summarization also can be domain-specific. Another possibility is using long-tail user efforts. Strategies such as crowd sourcing can be very helpful to make a large volume of labeled data.

7.2 Improve Scalability

The second important task for the better opinion summarization system is making the techniques scalable. The most useful and possible usage of opinion summarization is a web application. To satisfy users on the web, fast processing speed is necessary. However, so far, there has not been much consideration in scalability in this area. Because the volume of data on the web is huge and even keeps increasing, scalability is an unavoidable challenging. Moreover, using complicated NLP techniques for better accuracy makes scalability more challenging.

For better scalability, developing parallel algorithm is necessary. Cloud/grid computing is popularly used in web services because of its scalability. Developing parallelized algorithms for subtasks in opinion summarization is needed. Another way to make the service faster is using off-line processing as much as possible. Not all the computations are needed to be done on-line. As people designed index in information retrieval, well organized precomputed opinion indexes can make the system faster.

7.3 Need of Standardized Data Set and Evaluation Environment

Although there are some adhoc data sets for sentiment classification, there is no standardized data set for opinion summarization. Many of existing opinion summarization researchers crawled data from the web depending on their own purpose. Despite some of them published their data set [Hu and Liu 2004a; 2004b; 2006; Ku et al. 2006; Kim and Zhai 2009; Ganesan et al. 2010], there is no widely used data set yet.

Absence of evaluation measures which are widely used and cover entire opinion summarization steps is another issue to resolve. In previous researches, people usually evaluated the accuracy of the goal of each step and did not consider the entire step at once.

Good data sets and evaluation measures are necessary for fair comparison among different methods. Because each study uses different evaluation criteria on different data set, it is very difficult to judge whether one method is clearly better than the other or not. In addition, if the data set has publicized labeled data to evaluate, it would be even more helpful for researchers to try various methods or find the best one. Moreover, standardized data sets and measures will facilitate to start similar researches.

One possible way is following practice in general summarization area. In summarization, there are some evaluation measure such as ROUGE [Lin 2004] and pyramid method [Nenkova et al. 2007]. We may try to develop measures such as Opinion-ROUGE or Opinion-pyramid method which consider opinion aspects

7.4 Beyond Aspect-based Summarization

As we discussed above, many of opinion summarization works are in aspect-based summary format. Although aspect-based summarization encouraged developing techniques in each step, it also stereotyped opinion summarization. Non-three-step and non-aspect-based summary can provide novel benefits to users.

Integrated approaches may be useful even for better accuracy. Because each step is closely connected, using intermediate information from the previous and next steps can be useful for a more accuracy system. Like studies we introduced in Section 5, different types of summaries may provide novel aspects to users. For example, current summary cannot answer questions such as ‘what is the biggest complaint on the iPod screen’. Adopting Question/Answering techniques in opinion summarization also can be very useful.

7.5 Quality Control

Not all the opinions on the web are useful and trustable. Because there are many non-expert writers on the web, the quality of article varies a lot. Moreover, the more attention to opinionated articles, the more spamming attempts exist. Adapting quality control techniques is necessary for better opinion summarization.

As briefly covered by Liu’s tutorial [Liu 2008], spam detection techniques [Jindal and Liu 2008; Benczúr et al. 2005] can be useful for filtering out low quality articles. Opinion summarization also should pay attention to researches in trustworthiness, such as trust prediction/ranking [Agichtein et al. 2008; Kale et al. 2007; Liu et al. 2008; Juffinger et al. 2009] and expert finding [Zhang et al. 2007; Agarwal et al. 2008; Bouguessa et al. 2008]. Based on the obtained quality information, as opinion integration (see Section 5.2.1) does, handling articles depending on its quality can be another good way for high quality opinion summary.

ACKNOWLEDGMENTS

This survey is initiatied as a course project of CS512. We want to thank Prof. Jiawei Han for the precious comments.

REFERENCES

AGARWAL, N., LIU, H., TANG, L., AND YU, P. S. 2008. Identifying the influential bloggers in a community. In *WSDM '08: Proceedings of the international conference on Web search and web data mining*. ACM, New York, NY, USA, 207–218.

AGICHTEIN, E., CASTILLO, C., DONATO, D., GIONIS, A., AND MISHNE, G. 2008. Finding high-quality content in social media. In *WSDM '08: Proceedings of the international conference on Web search and web data mining*. ACM, New York, NY, USA, 183–194.

ARCHAK, N., GHOSE, A., AND IPEIROTIS, P. G. 2007. Show me the money!: deriving the pricing power of product features by mining consumer reviews. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, New York, NY, USA, 56–65.

BAGGA, A. AND BALDWIN, B. 1998. Algorithms for scoring coreference chains. In *In The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*. 563–566.

BALAHUR, A. AND MONTOYO, A. 2008. Multilingual feature-driven opinion extraction and summarization from customer reviews. In *NLDB '08: Proceedings of the 13th international conference on Natural Language and Information Systems*. Springer-Verlag, Berlin, Heidelberg, 345–346.

BENCZÚR, A. A., CSALOGÁNY, K., SARLÓS, T., AND UHER, M. 2005. Spamrank fully automatic link spam detection work in progress. In *Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web*.

BLEI, D. M., NG, A. Y., AND JORDAN, M. I. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res. 3*, 993–1022.

BOUGUESSA, M., DUMOULIN, B., AND WANG, S. 2008. Identifying authoritative actors in question-answering forums: the case of yahoo! answers. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, New York, NY, USA, 866–874.

CHEN, C., SANJUAN, F. I., SANJUAN, E., AND WEAVER, C. 2006. Visual analysis of conflicting opinions. Baltimore, MD, 35–42.

DAVE, K., LAWRENCE, S., AND PENNOCK, D. M. 2003. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*. ACM, New York, NY, USA, 519–528.

ETZIONI, O., CAFARELLA, M., DOWNEY, D., KOK, S., POPESCU, A.-M., SHAKED, T., SODERLAND, S., WELD, D. S., AND YATES, A. 2004. Web-scale information extraction in knowitall: (preliminary results). In *WWW '04: Proceedings of the 13th international conference on World Wide Web*. ACM, New York, NY, USA, 100–110.

GANESAN, K., ZHAI, C., AND HAN, J. 2010. Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics*. COLING '10. Association for Computational Linguistics, Stroudsburg, PA, USA, 340–348.

GANESAN, K., ZHAI, C., AND VIEGAS, E. 2012. Micropinion generation: an unsupervised approach to generating ultra-concise summaries of opinions. In *Proceedings of the 21st international conference on World Wide Web*. WWW '12. ACM, New York, NY, USA, 869–878.

HATZIVASSILOGLOU, V. AND MCKEOWN, K. R. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Morristown, NJ, USA, 174–181.

HOFMANN, T. 1999. Probabilistic latent semantic indexing. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, New York, NY, USA, 50–57.

HOVY, E. AND LIN, C.-Y. 1999. Automated text summarization in SUMMARIST. In *Advances in Automatic Text Summarization*, I. Mani and M. T. Maybury, Eds. MIT Press.

Hu, M. and Liu, B. 2004a. Mining and summarizing customer reviews. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, New York, NY, USA, 168–177.

Hu, M. and Liu, B. 2004b. Mining opinion features in customer reviews. In *AAAI'04: Proceedings of the 19th national conference on Artifical intelligence*. AAAI Press, 755–760.

Hu, M. and Liu, B. 2006. Opinion extraction and summarization on the web. In *AAAI'06: proceedings of the 21st national conference on Artificial intelligence*. AAAI Press, 1621–1624.

Hummel, R. A. and Zucker, S. W. 1987. On the foundations of relaxation labeling processes. 585–605.

J., M., Y., Z., Y., G., and H., Y. 1982. tong2yi4ci2ci2lin2. Shanghai Dictionary Press.

Jindal, N. and Liu, B. 2008. Opinion spam and analysis. In *WSDM '08: Proceedings of the international conference on Web search and web data mining*. ACM, New York, NY, USA, 219–230.

Juffinger, A., Granitzer, M., and Lex, E. 2009. Blog credibility ranking by exploiting verified content. In *WICOW '09: Proceedings of the 3rd workshop on Information credibility on the web*. ACM, New York, NY, USA, 51–58.

Kale, A., Karandikar, A., Kolari, P., Java, A., Finin, T., and Joshi, A. 2007. Modeling trust and influence in the blogosphere using link polarity. *International Conference on Weblogs and Social Media (ICWSM)*.

Kamps, J. and Marx, M. 2001. Words with attitude. In *In 1st International WordNet Conference*. 332–341.

Kim, H. D. and Zhai, C. 2009. Generating comparative summaries of contradictory opinions in text. In *CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management*. ACM, New York, NY, USA, 385–394.

Krippendorff, K. H. 2003. *Content Analysis: An Introduction to Its Methodology*, 2nd ed. Sage Publications, Inc.

Ku, L.-W., Liang, Y.-T., and Chen, H.-H. 2006. Opinion extraction, summarization and tracking in news and blog corpora. In *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*. 100–107.

Kupiec, J., Pedersen, J., and Chen, F. 1995. A trainable document summarizer. In *SIGIR '95: Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, New York, NY, USA, 68–73.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Proc. ACL workshop on Text Summarization Branches Out*. 10.

Lin, D. 1998. Dependency-based evaluation of minipar. In *Proc. Workshop on the Evaluation of Parsing Systems*. Granada.

Liu, B. 2006. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Liu, B. 2008. Opinion mining and summarization.

Liu, B. 2010. Sentiment analysis and subjectivity. In *Handbook of Natural Language Processing, Second Edition*, N. Indurkhya and F. J. Damerau, Eds. CRC Press, Taylor and Francis Group, Boca Raton, FL. ISBN 978-1420085921.

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

Liu, H., Lim, E.-P., Lauw, H. W., Le, M.-T., Sun, A., Srivastava, J., and Kim, Y. A. 2008. Predicting trusts among users of online communities: an epinions case study. In *EC '08: Proceedings of the 9th ACM conference on Electronic commerce*. ACM, New York, NY, USA, 310–319.

Lu, Y. and Zhai, C. 2008. Opinion integration through semi-supervised topic modeling. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*. ACM, New York, NY, USA, 121–130.

Lu, Y., Zhai, C., and Sundaresan, N. 2009. Rated aspect summarization of short comments. In *WWW '09: Proceedings of the 18th international conference on World wide web*. ACM, New York, NY, USA, 131–140.

Luo, X. 2005. On coreference resolution performance metrics. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Morristown, NJ, USA, 25–32.

MEI, Q., LING, X., WONDRA, M., SU, H., AND ZHAI, C. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*. ACM, New York, NY, USA, 171–180.

MISHNE, G. A., DE RIJKE, M., NICOLOV, N., SALVETTI, F., LIBERMAN, M., AND MARTIN, J. 2006. Moodviews: Tools for blog mood analysis. In *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW 2006)*. AAAI Press, AAAI Press, 153–154.

MULLEN, T. AND COLLIER, N. 2004. Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 412–418. Poster paper.

NENKOVA, A., PASSONNEAU, R., AND MCKEOWN, K. 2007. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Trans. Speech Lang. Process. 4,* 2, 4.

OSGOOD, C. E., SUCI, G. J., AND TANNENBAUM, P. H. 1967. *The Measurement of Meaning*. University of Illnois press, Urbana IL.

PAICE, C. D. 1990. Constructing literature abstracts by computer: techniques and prospects. *Inf. Process. Manage. 26,* 1, 171–186.

PANG, B. AND LEE, L. 2004. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, Morristown, NJ, USA, 271.

PANG, B. AND LEE, L. 2005. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, Morristown, NJ, USA, 115–124.

PANG, B. AND LEE, L. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr. 2,* 1-2, 1–135.

PANG, B., LEE, L., AND VAITHYANATHAN, S. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*. Association for Computational Linguistics, Morristown, NJ, USA, 79–86.

PASSONNEAU, R. J. 2004. Computing reliability for coreference annotation. In *In Proc. of LREC*. 1503–1506.

POPESCU, A.-M. AND ETZIONI, O. 2005. Extracting product features and opinions from reviews. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Morristown, NJ, USA, 339–346.

RILOFF, E. AND WIEBE, J. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*. Association for Computational Linguistics, Morristown, NJ, USA, 105–112.

RILOFF, E., WIEBE, J., AND WILSON, T. 2003. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*. Association for Computational Linguistics, Morristown, NJ, USA, 25–32.

ROSEN, S. 1974. Hedonic prices and implicit markets: Product differentiation in pure competition. *The Journal of Political Economy 82,* 1 (Jan-Feb), 34–55.

SNYDER, B. AND BARZILAY, R. 2007. Multiple aspect ranking using the good grief algorithm. In *In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL*. 300–307.

STOYANOV, V. AND CARDIE, C. 2006a. Partially supervised coreference resolution for opinion summarization through structured rule learning. In *EMNLP '06: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Morristown, NJ, USA, 336–344.

STOYANOV, V. AND CARDIE, C. 2006b. Toward opinion summarization: linking the sources. 9–14.

STOYANOV, V. AND CARDIE, C. 2008. Topic identification for fine-grained opinion analysis. In *COLING '08: Proceedings of the 22nd International Conference on Computational Linguistics*. Association for Computational Linguistics, Morristown, NJ, USA, 817–824.

TITOV, I. AND MCDONALD, R. 2008. Modeling online reviews with multi-grain topic models. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*. ACM, New York, NY, USA, 111–120.

TURNEY, P. D. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, Morristown, NJ, USA, 417–424.

TURNEY, P. D. AND LITTMAN, M. L. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Trans. Inf. Syst. 21,* 4, 315–346.

WIEBE, J. 2000. Learning subjective adjectives from corpora. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*. AAAI Press, 735–740.

WILSON, T., WIEBE, J., AND HWA, R. 2004. Just how mad are you? finding strong and weak opinion clauses. In *AAAI'04: Proceedings of the 19th national conference on Artifical intelligence*. AAAI Press, 761–767.

YU, H. AND HATZIVASSILOGLOU, V. 2003. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*. Association for Computational Linguistics, Morristown, NJ, USA, 129–136.

ZHAI, C., VELIVELLI, A., AND YU, B. 2004. A cross-collection mixture model for comparative text mining. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, New York, NY, USA, 743–748.

ZHANG, J., ACKERMAN, M. S., AND ADAMIC, L. 2007. Expertise networks in online communities: structure and algorithms. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*. ACM, New York, NY, USA, 221–230.

ZHUANG, L., JING, F., AND ZHU, X.-Y. 2006. Movie review mining and summarization. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*. ACM, New York, NY, USA, 43–50.