Review Sentiment Scoring via a Parse-and-Paraphrase Paradigm

Jingjing Liu, Stephanie Seneff
MIT Computer Science & Artificial Intelligence Laboratory
32 Vassar Street, Cambridge, MA 02139
{jingl, seneff}@csail.mit.edu

Abstract

This paper presents a parse-and-paraphrase paradigm to assess the degrees of sentiment for product reviews. Sentiment identification has been well studied; however, most previous work provides binary polarities only (positive and negative), and the polarity of sentiment is simply reversed when a negation is detected. The extraction of lexical features such as unigram/bigram also complicates the sentiment classification task, as linguistic structure such as implicit long-distance dependency is often disregarded. In this paper, we propose an approach to extracting adverb-adjective-noun phrases based on clause structure obtained by parsing sentences into a hierarchical representation. We also propose a robust general solution for modeling the contribution of adverbials and negation to the score for degree of sentiment. In an application involving extracting aspect-based pros and cons from restaurant reviews, we obtained a 45% relative improvement in recall through the use of parsing methods, while also improving precision.

1 Introduction

Online product reviews have provided an extensive collection of free-style texts as well as product ratings prepared by general users, which in return provide grassroots contributions to users interested in a particular product or service as assistance. Yet, valuable as they are, free-style reviews contain much noisy data and are tedious to read through in order to reach an overall conclusion. Thus, we conducted this study to automatically process and evaluate product reviews in order to generate both numerical evaluation and textual summaries of users' opinions, with the ultimate goal of adding value to real systems such as a restaurant-guide dialogue system.

Sentiment summarization has been well studied in the past decade (Turney, 2002; Pang et al., 2002; Dave et al., 2003; Hu and Liu, 2004a, 2004b; Carenini et al., 2006; Liu et al., 2007). The polarity of users' sentiments in each segment of review texts is extracted, and the polarities of individual sentiments are aggregated among all the sentences/segments of texts to give a numerical scaling on sentiment orientation.

Most of the work done for sentiment analysis so far has employed shallow parsing features such as part-of-speech tagging. Frequent adjectives and nouns/noun phrases are extracted as opinion words and representative product features. However, the linguistic structure of the sentence is usually not taken into consideration. High level linguistic features, if well utilized and accurately extracted, can provide much insight into the semantic meaning of user opinions and contribute to the task of sentiment identification.

Furthermore, in addition to adjectives and nouns, adverbials and negation also play an important role in determining the degree of the orientation level. For example, "very good" and "good" certainly express different degrees of positive sentiment. Also, in previous studies, when negative expressions are identified, the polarity of sentiment in the associated segment of text is simply reversed. However, semantic expressions are quite different from the absolute opposite values in mathematics. For example, "not bad" does not express the opposite meaning of "bad", which would be highly positive. Simply reversing the polarity of sentiment on the appearance of negations may result in inaccurate interpretation of sentiment expressions. Thus, a system which attempts to quantify sentiment while ignoring adverbials is missing a significant component of the sentiment score, especially if the adverbial is a negative word.

Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 161–169,
Singapore, 6-7 August 2009.

Another challenging aspect of negation is proper scoping of the negative reference over the right constituent, which we argue, can be handled quite well with careful linguistic analysis. Take the sentence "*I don't think the place is very clean*" as example. A linguistic approach associating long-distance elements with semantic relations can identify that the negation "not" scopes over the complement clause, thus extracting "not very clean" instead of "very clean".

Our goal in modeling adverbials is to investigate whether a simple linear correction model can capture the polarity contribution of all adverbials. Furthermore, is it also appropriate to adjust for multiple adverbs, including negation, via a linear *additive* model? I.e., can "not very good" be modeled as *not(very(good))*? The fact that "not very good" seems to be *less* negative than "not good" suggests that such an algorithm might work well. From these derivations we have developed a model which treats negations in the exact same way as modifying adverbs, via an accumulative linear offset model. This yields a very generic and straightforward solution to modeling the strength of sentiment expression.

In this paper we utilize a parse-and-paraphrase paradigm to identify semantically related phrases in review texts, taking quantifiers (e.g., modifying adverbs) and qualifiers (e.g., negations) into special consideration. The approach makes use of a lexicalized probabilistic syntactic grammar to identify and extract sets of *adverb-adjective-noun* phrases that match review-related patterns. Such patterns are constructed based on well-formed linguistic structure; thus, relevant phrases can be extracted reliably.

We also propose a cumulative linear offset model to calculate the degree of sentiment for joint adjectives and quantifiers/qualifiers. The proposed sentiment prediction model takes modifying adverbs and negations as universal scales on strength of sentiment, and conducts cumulative calculation on the degree of sentiment for the associated adjective. With this model, we can provide not only qualitative textual summarization such as "good food" and "bad service", but also a numerical scoring of sentiment, i.e., "how good the food is" and "how bad the service is."

2 Related Work

There have been many studies on sentiment classification and opinion summarization (Pang and Lee, 2004, 2005; Gamon et al., 2005; Popescu and Etzioni, 2005; Liu et al., 2005; Zhuang et al., 2006; Kim and Hovy, 2006). Specifically, aspect rating as an interesting topic has also been widely studied (Titov and McDonald, 2008a; Snyder and Barzilay, 2007; Goldberg and Zhu, 2006). Recently, Baccianella et. al. (2009) conducted a study on multi-facet rating of product reviews with special emphasis on how to generate vectorial representations of the text by means of POS tagging, sentiment analysis, and feature selection for ordinal regression learning. Titov and McDonald (2008b) proposed a joint model of text and aspect ratings which utilizes a modified LDA topic model to build topics that are representative of ratable aspects, and builds a set of sentiment predictors. Branavan et al. (2008) proposed a method for leveraging unstructured annotations in product reviews to infer semantic document properties, by clustering user annotations into semantic properties and tying the induced clusters to hidden topics in the text.

3 System Overview

Our review summarization task is to extract sets of descriptor-topic pairs (e.g., "excellent service") from a set of reviews (e.g., for a particular restaurant), and to cluster the extracted phrases into representative aspects on a set of dimensions (e.g., "food", "service" and "atmosphere"). Driven by this motivation, we propose a three-stage system that automatically processes reviews. A block diagram is given in Figure 1.

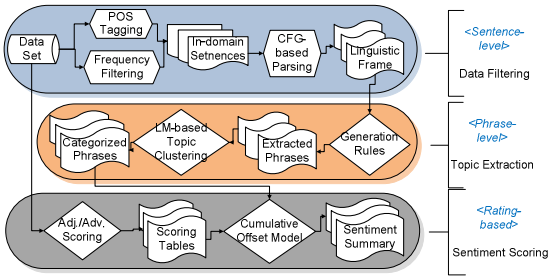

Figure 1. Framework of review processing.

The first stage is *sentence-level data filtering*. Review data published by general users is often in free-style, and a large fraction of the data is either ill-formed or not relevant to the task. We classify these as *out of domain* sentences. To filter out such noisy data, we collect unigram statistics on all the relevant words in the corpus, and select high frequency adjectives and nouns. Any sentence that contains none of the high-frequency nouns or adjectives is rejected from further analysis. The remaining *in-domain* sentences are subjected to the second stage, *parse*

analysis and semantic understanding, for topic extraction.

From the parsable sentences we extract descriptor-topic phrase patterns based on a carefully-designed generation grammar. We then apply *LM* (*language model) based topic clustering* to group the extracted phrases into representative aspects. The third stage scores the *degree of sentiment* for adjectives, as well as the *strength of sentiment* for modifying adverbs and negations, which further refine the degree of sentiment of the associated adjectives. We then run a *linear additive model* to assign a combined sentiment score for each extracted phrase.

The rest of the paper is structured as follows: In Section 4, we explain the linguistic analysis. In Section 5, we describe the cumulative model for assessing the degree of sentiment. Section 6 provides a systematic evaluation, conducted on real data in the restaurant review domain harvested from the Web. Section 7 provides a discussion analyzing the results. Section 8 summarizes the paper as well as pointing to future work.

4 Linguistic Analysis

4.1 Parse-and-Paraphrase

Our linguistic analysis is based on a parse-and-paraphrase paradigm. Instead of the flat structure of a surface string, the parser provides a hierarchical representation, which we call a *linguistic frame* (Xu et al., 2008). It preserves linguistic structure by encoding different layers of semantic dependencies. The grammar captures syntactic structure through a set of carefully constructed context free grammar rules, and employs a feature-passing mechanism to enforce long distance constraints. The grammar is lexicalized, and uses a statistical model to rank order competing hypotheses. It knows explicitly about 9,000 words, with all unknown words being interpreted as nouns. The grammar probability model was trained automatically on the corpus of review sentences.

To produce the phrases, a set of generation rules is carefully constructed to only extract sets of related adverbs, adjectives and nouns. The adjective-noun relationships are captured from the following linguistic patterns: (1) all adjectives attached directly to a noun in a noun phrase, (2) adjectives embedded in a relative clause modifying a noun, and (3) adjectives related to nouns in a subject-predicate relationship in a clause. These patterns are compatible, i.e., if a clause contains both a modifying adjective and a predicate adjective related to the same noun, two adjective-noun pairs are generated by different patterns. As in, "The efficient waitress was nonetheless very courteous." It is a "parse-and-paraphrase-like" paradigm: the paraphrase tries to preserve the original words intact, while reordering them and/or duplicating them into multiple NP units. Since they are based on syntactic structure, the generation rules can also be applied in any other domain involving opinion mining.

An example linguistic frame is shown in Figure 2, which encodes the sentence "*The caesar with salmon or chicken is really quite good.*" In this example, for the adjective "good", the nearby noun "chicken" would be associated with it if only proximity is considered. From the linguistic frame, however, we can easily associate "caesar" with "good" by extracting the head of the topic sub-frame and the head of the predicate sub-frame, which are encoded in the same layer (root layer) of the linguistic frame. Also, we can tell from the predicate sub-frame that there is an adverb "*quite*" modifying the head word "*good*". The linguistic frame also encodes an adverb "*really*" in the upstairs layer. A well-constructed generation grammar can create customized *adverb-adjective-noun* phrases such as "quite good caesar" or "really quite good caesar".

```
{c cstatement
 :topic {q caesar
         :quantifier "def"
         :pred {p with :topic {q salmon
                               :pred {p conjunction
                                      :or {q chicken }}}}
 :adv "really"
 :pred {p adj_complement
        :pred {p adjective
               :adv "quite"
               :pred {p quality :topic "good" }}}}
```

Figure 2. Linguistic frame for "*The caesar with salmon or chicken is really quite good.*"

Interpreting negation in English is not straightforward, and it is often impossible to do correctly without a deep linguistic analysis. Xuehui Wu (2005) wrote: "The scope of negation is a complex linguistic phenomenon. It is easy to perceive but hard to be defined from a syntactic point of view. Misunderstanding or ambiguity may occur when the negative scope is not understood clearly and correctly." The majority rule for negation is that it scopes over the remainder of its containing clause, and this works well for most cases. For example, Figure 3 shows

the linguistic frame for the sentence "*Their menu was a good one that didn't try to do too much.*"

```
{c cstatement
 :topic {q menu   :poss "their" } }
 :complement {q pronoun   :name "one"
        :adj_clause {c cstatement
                :conjn "that"
                :negate "not"
                :pred {p try :to_clause  {p do
                        :topic {q object
                        :adv "too"
                        :quant "much" }}}}
 :pred {p adjective
        :pred {p quality :topic "good" }}}
```

Figure 3. Linguistic frame for "*Their menu was a good one that didn't try to do too much.*"

Traditional approaches which do not consider the linguistic structure would treat the appearance of "*not*" as a negation and simply reverse the sentiment of the sentence to negative polarity, which is wrong as the sentence actually expresses positive opinion for the topic "menu". In our approach, the negation "*not*" is identified as under the sub-frame of the complement clause, instead of in the same or higher layer of the adjective sub-frame; thus it is considered as unrelated to the adjective "good". In this way we can successfully predict the scope of the reference of the negation over the correct constituent of a sentence and create proper association between negation and its modified words.

4.2 LM-based Topic Clustering

To categorize the extracted phrases into representative aspects, we automatically group the identified topics into a set of clusters based on LM probabilities. The LM-based algorithm assumes that topics which are semantically related have high probability of co-occurring with similar descriptive words. For example, "delicious" might co-occur frequently with both "pizza" and "dessert". By examining the distribution of bigram probability of these topics with corresponding descriptive words, we can group "pizza" and "dessert" into the same cluster of "food".

We select a small set of the most common topics, i.e., topics with the highest frequency counts, and put them into an initial set *I*. Then, for each candidate topic t_c outside set *I*, we calculate its probability given each topic t_i within the initial set *I*, given by:

$$P(t_c|\, t_i) = \sum_{a\in A} P(t_c|a) \cdot P(a|t_i)$$
$$= \sum_{a\in A} \frac{P(a,t_c)}{P(a)} \cdot \frac{P(a,t_i)}{P(t_i)}$$
$$= \frac{1}{P(t_i)} \sum_{a\in A} \frac{1}{P(a)} \cdot P(a,t_c) \cdot P(a,t_i) \quad (1)$$

where *A* represents the set of all the adjectives in the corpus. For each candidate topic t_c, we choose the cluster of the initial topic t_i with which it has the highest probability score.

There are also cases where a meaningful adjective occurs in the absence of an associated topic, e.g., "It is quite *expensive*." We call such cases the "*widow-adjective*" case. Without hard-coded ontology matching, it is difficult to identify "expensive" as a price-related expression. To discover such cases and associate them with related topics, we propose a "*surrogate topic*" matching approach based on bigram probability.

As aforementioned, the linguistic frame organizes all adjectives into separate clauses. Thus, we create a "surrogate topic" category in the linguistic frames for widow-adjective cases, which makes it easy to detect descriptors that are affiliated with uninformative topics like the pronoun "it". We then have it generate phrases such as "expensive surrogate_topic" and use bigram probability statistics to automatically map each sufficiently strongly associated adjective to its most common topic among our major classes, e.g., mapping "expensive" with its surrogate topic "price". Therefore, we can generate sets of additional phrases in which the topic is "hallucinated" from the widow-adjective.

5 Assessment of Sentiment Strength

5.1 Problem Formulation

Given the sets of adverb-adjective-noun phrases extracted by linguistic analysis, our goal is to assign a score for the degree of sentiment to each phrase and calculate an average rating for each aspect. An example summary is given in Table 1.

Table 1. Example of review summary.

Aspect	Extracted phrases	Rating
Atmosphere	very nice ambiance, outdoor patio	4.8
Food	not bad meal, quite authentic food	4.1
Place	not great place, very smoky restaurant	2.8
Price	so high bill, high cost, not cheap price	2.2

To calculate the numerical degree of sentiment, there are three major problems to solve: 1) how to associate *numerical scores* with *textual sentiment*; 2) whether to calculate sentiment scores for adjectives and adverbs *jointly* or *separately*; 3)

whether to treat negations as *special cases* or in the *same way* as modifying adverbs.

There have been studies on building sentiment lexicons to define the strength of sentiment of words. Esuli and Sebastiani (2006) constructed a lexical resource, SentiWordNet, a WordNet-like lexicon emphasizing sentiment orientation of words and providing numerical scores of how objective, positive and negative these words are. However, lexicon-based methods can be tedious and inefficient and may not be accurate due to the complex cross-relations in dictionaries like WordNet. Instead, our primary approach to sentiment scoring is to make use of collective data such as user ratings. In product reviews collected from online forums, the format of a review entry often consists of three parts: pros/cons, free-style text and user rating. We assume that user rating is normally consistent with the tone of the review text published by the same user. By associating user ratings with each phrase extracted from review texts, we can easily associate numerical scores with textual sentiment.

A simple strategy of rating assignment is to take each extracted adverb-adjective pair as a composite unit. However, this method is likely to lead to a large number of rare combinations, thus suffering from sparse data problems. Therefore, an interesting question to ask is whether it is feasible to assign to each adverb a perturbation score, which adjusts the score of the associated adjective up or down by a fixed scalar value. This approach thus hypothesizes that "very expensive" is as much worse than "expensive" as "very romantic" is better than "romantic". This allows us to pool all instances of a given adverb regardless of which adjective it is associated with, in order to compute the absolute value of the perturbation score for that adverb. Therefore, we consider adverbs and adjectives separately when calculating the sentiment score, treating each modifying adverb as a universal quantifier which consistently scales up/down the strength of sentiment for the adjectives it modifies.

Furthermore, instead of treating negation as a special case, the universal model works for all adverbials. The model hypothesizes that "not bad" is as much better than "bad" as "not good" is worse than "good", i.e., negations push positive/negative adjectives to the other side of sentiment polarity by a *universal scale*. This again, allows us to pool all instances of a given negation and compute the absolute value of the perturbation score for that negation, in the same way as dealing with modifying adverbs.

5.2 Linear Additive Model

For each adjective, we average all its ratings given by:

$$Score(adj) = \frac{\sum_{i \in P} \frac{N}{n_{r_i}} \cdot r_i}{\sum_{r_i} \frac{N}{n_{r_i}}} \qquad (2)$$

where P represents the set of appearances of adjective adj, r_i represents the associated user rating in each appearance of adj, N represents the number of entities (e.g., restaurants) in the entire data set, and n_{r_i} represents the number of entities with rating r_i. The score is averaged over all the appearances, weighted by the frequency count of each category of rating to remove bias towards any category.

As for adverbs, using a slightly modified version of equation (2), we can get a rating table for all *adverb-adjective* pairs. For each adverb *adv*, we get a list of all its possible combinations with adjectives. Then, for each *adj* in the list, we calculate the distance between the rating of *adv-adj* and the rating of the *adj* alone. We then aggregate the distances among all the pairs of *adv-adj* and *adj* in the list, weighted by the frequency count of each *adv-adj* pair:

$$Score(adv) = \sum_{t \in A} \frac{count(adv, adj_t)}{\sum_{j \in A} count(adv, adj_j)} \cdot Pol(adj_t) \cdot (r(adv, adj_t) - r(adj_t)) \qquad (3)$$

where $count(adv,\ adj_t)$ represents the count of the combination $adv - adj_t$, A represents the set of adjectives that co-occur with adv, $r(adv, adj_t)$ represents the sentiment rating of the combination $adv - adj_t$, and $r(adj_t)$ represents the sentiment rating of the adjective adj_t alone. $Pol(adj_t)$ represents the polarity of adj_t, assigned as *1* if adj_t is positive, and *-1* if negative.

Specifically, negations are well handled by the same scoring strategy, treated exactly the same as modifying adverbs, except that they get such strong negative scores that the sentiment of the associated adjectives is pushed to the other side of the polarity scale.

After obtaining the strength rating for adverbs and the sentiment rating for adjectives, the next step is to assign the strength of sentiment to each phrase (negation-adverb-adjective-noun) extracted by linguistic analysis, as given by:

$$Score\left(neg\left(adv(adj)\right)\right) = r(adj) + Pol(adj) \cdot r(adv) + Pol(adj) \cdot r(neg) \qquad (4)$$

where $r(adj)$ represents the rating of adjective adj, $r(adv)$ represents the rating of adverb adv, and $r(neg)$ represents the rating of negation neg. $Pol(adj)$ represents the polarity of adj, assigned as *1* if adj is positive, and *-1* if negative. Thus, if adj is positive, we assign a combined rating $r(adj) + r(adv)$ to this phrase. If it is negative, we assign $r(adj) - r(adv)$. Specifically, if it is a negation case, we further assign a linear offset $r(neg)$ if adj is positive or $-r(neg)$ if adj is negative. For example, given the ratings <good: 4.5>, <bad: 1.5>, <very: 0.5> and <not: -3.0>, we would assign "5.0" to "very good" (*score(very(good))=4.5+0.5*), "1.0" to "very bad" (*score(very(bad))=1.5-0.5*), and "2.0" to "not very good" (s*core(not(very(good)))= 4.5+0.5-3.0*). The corresponding sequence of different degrees of sentiment is: "*very good*: 5.0" > "*good*: 4.5" > "*not very good*: 2.0" > "*bad*: 1.5" > "*very bad*: 1.0".

6 Experiments

In this section we present a systematic evaluation of the proposed approaches conducted on real data. We crawled a data collection of 137,569 reviews on 24,043 restaurants in 9 cities in the U.S. from an online restaurant evaluation website[1]. Most of the reviews have both pros/cons and free-style text. For the purpose of evaluation, we take those reviews containing pros/cons as the experimental set, which is 72.7% (99,147 reviews) of the original set.

6.1 Topic Extraction

Based on the experimental set, we first filtered out-of-domain sentences based on frequency count, leaving a set of 857,466 in-domain sentences (67.5%). This set was then subjected to parse analysis; 78.6% of them are parsable.

Given the parsing results in the format of linguistic frame, we used a set of language generation rules to extract relevant adverb-adjective-noun phrases. We then selected the most frequent 6 topics that represented appropriate dimensions for the restaurant domain ("place", "food", "service", "price", "atmosphere" and "portion") as the initial set, and clustered the extracted topic mentions into different aspect categories by creating a set of topic mappings with the LM-based clustering method. Phrases not belonging to any category are filtered out.

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

To evaluate the performance of the proposed approach (LING) to topic extraction, we compare it with a baseline method similar to (Hu and Liu, 2004a, 2004b; Liu et al., 2005). We performed part-of-speech tagging on both parsable and unparsable sentences, extracted each pair of noun and adjective that has the smallest proximity, and filtered out those with low frequency counts. Adverbs and negation words that are adjacent to the identified adjectives were also extracted along with the adjective-noun pairs. We call this the "neighbor baseline" (NB).

The proposed method is unable to make use of the non-parsable sentences, which make up over 20% of the data. Hence, it seems plausible to utilize a back-off mechanism for these sentences via a combined system (COMB) incorporating NB only for the sentences that fail to parse.

In considering how to construct the "ground truth" set of pros and cons for particular aspects, our goal was to minimize error as much as possible without requiring exorbitant amounts of manual labeling. We also wanted to assure that the methods were equally fair to both systems (LING and NB). To these ends, we decided to pool together all of the topic mappings and surrogate topic hallucinations obtained automatically from both systems, and then to manually edit the resulting list to eliminate any that were deemed unreasonable. We then applied these edited mappings in an automatic procedure to the adjective-noun pairs in the user-provided pros and cons of all the restaurant reviews. The resulting aspect-categorized phrase lists are taken as the ground truth. Each system then used its own (unedited) set of mappings in processing the associated review texts.

We also needed an algorithm to decide on a particular set of reviews for consideration, again, with the goal of omitting bias towards either of the two systems. We decided to retain as the evaluation set all reviews which obtained at least one topic extraction from both systems. Thus the two systems processed exactly the same data with exactly the same definitions of "ground truth". Performance was evaluated on this set of 62,588 reviews in terms of recall (percentage of topics in the ground truth that are also identified from the review body) and precision (percentage of extracted topics that are also in the ground truth). These measures are computed separately for each review, and then averaged over all reviews.

As shown in Table 2, without clustering, the LING approach gets 4.6% higher recall than the

NB baseline. And the recall from the COMB approach is 3.9% higher than that from the LING approach and 8.5% higher than that from the NB baseline. With topic clustering, the COMB approach also gets the highest recall, with a 4.9% and 17.5% increase from the LING approach and the NB baseline respectively. The precision is quite close among the different approaches, around 60%. Table 2 also shows that the topic clustering approach increases the recall by 4.8% for the NB baseline, 12.8% for the LING approach, and 13.8% for the COMB approach.

Table 2. Experimental results of topic extraction by the NB baseline, the proposed LING approach and a combined system (COMB).

	No Clustering		
	NB	**LING**	**COMB**
Recall	39.6%	44.2%	48.1%
Precision	60.2%	60.0%	59.8%
	With Clustering		
	NB	**LING**	**COMB**
Recall	44.4%	57.0%	61.9%
Precision	56.8%	61.1%	60.8%

6.2 Sentiment Scoring

To score the degree of sentiment for each extracted phrase, we built a table of sentiment score (<*adjective*: *score*>) for adjectives and a table of strength score (<*adverb*: *score*>) for adverbs. The pros/cons often contain short and well-structured phrases, and have better parsing quality than the long and complex sentences in free-style texts; pros/cons also have clear sentiment orientations. Thus, we use pros/cons to score the sentiment of adjectives, which requires strong polarity association. To obtain reliable ratings, we associate the adjectives in the "*pros*" of review entries which have a user rating 4 or 5, and associate the adjectives in the "*cons*" of review entries with user ratings 1 or 2 (the scale of user rating is 1 to 5). Reviews with rating 3 are on the boundary of sentiment, so we associate both sides with the overall rating. On the other hand, the frequencies of adverbs in free-style texts are much higher than those in pros/cons, as pros/cons mostly contain *adjective-noun* patterns. Thus, we use free-style texts instead of pros/cons to score the strength of adverbs.

Partial results of the sentiment scoring are shown in Tables 3 and 4. As shown in Table 3, the polarity of sentiment as well as the degree of polarity of an adjective can be distinguished by its score. The higher the sentiment score is, the more positive the adjective is.

Table 3. Sentiment scoring for selected adjectives.

Adjective	**Rating**	**Adjective**	**Rating**
Excellent	5.0	Awesome	4.8
Easy	4.1	Great	4.4
Good	3.9	Limited	3.4
Inattentive	2.75	Overpriced	2.3
Rude	1.69	Horrible	1.3

Table 4 gives the scores of strength for most common adverbs. The higher the strength score is, the more the adverb scales up/down the degree of sentiment of the adjective it modifies. While "not" gets a strong negative score, some adverbs such as "a little" (-0.65) and "a bit" (-0.83) also get negative scores, indicating slightly less sentiment for the associated adjectives.

Table 4. Strength scoring for selected adverbs.

Adverb	**Rating**	**Adverb**	**Rating**
Super	0.58	Fairly	0.13
Extremely	0.54	Pretty	0.07
Incredibly	0.49	A little	-0.65
Very	0.44	A bit	-0.83
Really	0.39	Not	-3.10

To evaluate the performance of sentiment scoring, we randomly selected a subset of 1,000 adjective-noun phrases and asked two annotators to independently rate the sentiment of each phrase on a scale of 1 to 5. We compared the sentiment scoring between our system and the annotations in a measurement of mean distance:

$$distance = \frac{1}{|S|}\sum_{p \in S} |r_{ip} - r_{ap}| \quad (5)$$

where S represents the set of phrases, p represents each phrase in the set S, r_{ip} represents the rating on phrase p from our sentiment scoring system, and r_{ap} represents the annotated rating on phrase p. As shown in Table 5, the obtained mean distance between the scoring from our approach and that from each annotation set is 0.46 and 0.43 respectively, based on the absolute rating scale from 1 to 5. This shows that the scoring of sentiment from our system is quite close to human annotation. The kappa agreement between the two annotation sets is 0.68, indicating high consistency between the annotators. The reliability of these results gives us sufficient confidence to make use of the scores of sentiments for summarization.

To examine the prediction of sentiment polarity, for each annotation set, we pooled the phrases with rating 4/5 into "positive", rating 1/2 into "negative", and rating 3 into "neutral". Then we rounded up the sentiment scores from our system to integers and pooled the scores into three polar-

ity sets ("positive", "negative" and "neutral") in the same way. As shown in Table 5, the obtained kappa agreement between the result from our system and that from each annotation set is 0.55 and 0.60 respectively. This shows reasonably high agreement on the polarity of sentiment between our system and human evaluation.

Table 5. Comparison of sentiment scoring between the proposed approach and two annotation sets.

	Annotation 1	**Annotation 2**
Mean distance	0.46	0.43
Kappa agreement	0.55	0.60

Table 6. Experimental results of topic extraction based on sentiment polarity matching.

	No Clustering		
	NB	**LING**	**COMB**
Recall	34.5%	38.9%	42.2%
Precision	53.8%	54.0%	53.3%
	With Clustering		
	NB	**LING**	**COMB**
Recall	37.4%	49.7%	54.1%
Precision	48.5%	52.9%	51.4%

To evaluate the combination of topic extraction and sentiment identification, we repeated the topic extraction experiments presented in Table 2, but this time requiring as well a correct polarity assignment to obtain a match with the pros/cons ground truth. As shown in Table 6, the COMB approach gets the highest recall both with and without topic clustering, and the recall from the LING approach is higher than that from the NB baseline in both cases as well, indicating the superiority of the proposed approach. The precision is stable among the different approaches, consistent with the case without the consideration of sentiment polarity.

7 Discussion

It is surprising that the parse-and-paraphrase method performs so well, despite the fact that it utilizes less than 80% of the data (parsable set). In this section, we will discuss two experiments that were done to tease apart the contributions of different variables. In both experiments, we compared the change in relative improvement in recall between NB and LING, relative to the values in Table 6, in the with-clustering condition. In the table, LING obtains a score of 49.7% for recall, which is a 33% relative increase from the score for NB (37.4%). Three distinct factors could play a role in the improvement: the widow-adjective topic hallucinations, the topic mapping for clustering, and the extracted phrases themselves. An experiment involving omitting topic hallucinations from widow adjectives determined that these account for 12% of the relative increase. To evaluate the contribution of clustering, we replaced the mapping tables used by both systems with the edited one used by the ground truth computation. Thus, both systems made use of the same mapping table, removing this variable from consideration. This improved the performance of both systems (NB and LING), but resulted in a decrease of LING's relative improvement by 17%. This implies that LING's mapping table is superior. Since both systems use the same sentiment scores for adjectives and adverbs, the remainder of the difference (71%) must be due simply to higher quality extracted phrases.

We suspected that over-generated phrases (the 40% of phrases that find no mappings in the pros/cons) might not really be a problem. To test this hypothesis, we selected 100 reviews for their high density of extracted phrases, and manually evaluated all the over-generated phrases. We found that over 80% were well formed, correct, and informative. Therefore, a lower precision here does not necessarily mean poor performance, but instead shows that the pros/cons provided by users are often incomplete. By extracting summaries from review texts we can recover additional valuable information.

8 Conclusions & Future Work

This paper presents a parse-and-paraphrase approach to assessing the degree of sentiment for product reviews. A general purpose context free grammar is employed to parse review sentences, and semantic understanding methods are developed to extract representative negation-adverb-adjective-noun phrases based on well-defined semantic rules. A language modeling-based method is proposed to cluster topics into respective categories. We also introduced in this paper a cumulative linear offset model for supporting the assessment of the strength of sentiment in adjectives and quantifiers/qualifiers (including negations) on a numerical scale. We demonstrated that the parse-and-paraphrase method can perform substantially better than a neighbor baseline on topic extraction from reviews even with less data. The future work focuses in two directions: (1) building a relational database from the summaries and ratings and using it to enhance users' experiences in a multimodal spoken dialogue system; and (2) applying our techniques to other domains to demonstrate generality.

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2009. Multi-facet Rating of Product Reviews. In Proceedings of European Conference on Information Retrieval.

S.R.K. Branavan, Harr Chen, Jacob Eisenstein, and Regina Barzilay. 2008. Learning document-level semantic properties from free-text annotations. In Proceedings of the Annual Conference of the Association for Computational Linguistics.

Giuseppe Carenini, Raymond Ng, and Adam Pauls 2006. Multi-Document Summarization of Evaluative Text. In Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics.

Kushal Dave, Steve Lawrence, and David M. Pennock. 2003. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In Proceedings of the International Conference on World Wide Web.

Andrea Esuli and Fabrizio Sebastiani. 2006. SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In Proceedings of the 5th Conference on Language Resources and Evaluation.

Michael Gamon, Anthony Aue, Simon Corston-Oliver, and Eric Ringger . 2005. Pulse: Mining customer opinions from free text. In Proceedings of the 6th International Symposium on Intelligent Data Analysis.

Andrew Goldberg and Xiaojin Zhu. 2006. Seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization. In HLT-NAACL 2006 Workshop on Textgraphs: Graph-based Algorithms for Natural Language Processing.

Minqing Hu and Bing Liu. 2004a. Mining and summarizing customer reviews. In Proceedings of the 2004 ACM SIGKDD international conference on Knowledge Discovery and Data mining.

Minqing Hu and Bing Liu. 2004b. Mining Opinion Features in Customer Reviews. In Proceedings of Nineteenth National Conference on Artificial Intelligence.

Soo-Min Kim and Eduard Hovy. 2006. Automatic identification of pro and con reasons in online reviews. In Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 483–490.

Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion observer: analyzing and comparing opinions on the web. In Proceedings of International Conference on World Wide Web.

Jingjing Liu, Yunbo Cao, Chin-Yew Lin, Yalou Huang, and Ming Zhou. 2007. Low-Quality Product Review Detection in Opinion Summarization. In Proceedings of the Conference on Empirical Methods in Natural Language Processing.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the Annual Conference of the Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In Proceedings of the Annual Conference of the Association for Computational Linguistics.

Bo Pang, Lillian Lee, and S. Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing.

A.M. Popescu and O. Etzioni. 2005. Extracting product features and opinions from reviews. In Proceedings of the Conference on Empirical Methods in Natural Language Processing.

Benjamin Snyder and Regina Barzilay. 2007. Multiple Aspect Ranking using the Good Grief Algorithm. In Proceedings of the Joint Conference of the North American Chapter of the Association for Computational Linguistics and Human Language Technologies.

Ivan Titov and Ryan McDonald. 2008a. Modeling online reviews with multi-grain topic models. In Proceedings of the 17h International Conference on World Wide Web.

Ivan Titov and Ryan McDonald. 2008b. A Joint Model of Text and Aspect Ratings for Sentiment Summarization. In Proceedings of the Annual Conference of the Association for Computational Linguistics.

Peter D. Turney. 2002. Thumbs up or thumbs down? Sentiment orientation applied to unsupervised classification of reviews. In Proceedings of the Annual Conference of the Association for Computational Linguistics.

Xuehui Wu, 2005. On the Scope of Negation in English, Sino-US English Teaching, Vol. 2, No. 9, Sep. 2005. pp. 53-56.

Yushi Xu, Jingjing Liu, Stephanie Seneff. 2008. Mandarin Language Understanding in Dialogue Context. In Proceedings of International Symposium on Chinese Spoken Language Processing.

Li Zhuang, Feng Jing, and Xiao-Yan Zhu. 2006. Movie review mining and summarization. In Proceedings of the 15th ACM international conference on Information and knowledge management.