A Hybrid Approach to Multi-document Summarization of Opinions in Reviews

Giuseppe Di Fabbrizio
Amazon.com*
Cambridge, MA - USA
pino@difabbrizio.com

Amanda J. Stent
Yahoo! Labs
New York, NY - USA
stent@labs.yahoo.com

Robert Gaizauskas
Department of Computer Science
University of Sheffield, Sheffield - UK
R.Gaizauskas@sheffield.ac.uk

Abstract

We present a hybrid method to generate summaries of product and services reviews by combining natural language generation and salient sentence selection techniques. Our system, STARLET-H, receives as input textual reviews with associated rated topics, and produces as output a natural language document summarizing the opinions expressed in the reviews. STARLET-H operates as a hybrid abstractive/extractive summarizer: using extractive summarization techniques, it selects salient quotes from the input reviews and embeds them into an automatically generated abstractive summary to provide evidence for, exemplify or justify positive or negative opinions. We demonstrate that, compared to extractive methods, summaries generated with abstractive and hybrid summarization approaches are more readable and compact.

1 Introduction

Text summarization is a well-established area of research. Many approaches are *extractive*, that is, they select and stitch together pieces of text from the input documents (Goldstein et al., 2000; Radev et al., 2004). Other approaches are *abstractive*; they use natural language generation (NLG) techniques to paraphrase and condense the content of the input documents (Radev and McKeown, 1998). Most summarization methods focus on distilling *factual* information by identifying the input documents' main topics, removing redundancies, and coherently ordering extracted phrases or sentences. Summarization of *sentiment-laden text* (e.g., product or service reviews) is substantially different from the traditional text summarization task: instead of presenting facts, the summarizer must present the range of opinions and the consensus opinion (if any), and instead of focusing on one topic, the summarizer must present information about multiple aspects of the target entity. In addition, traditional summarization techniques discard redundancies, while for summarization of sentiment-laden text, similar opinions mentioned multiple times across documents are crucial indicators of the overall strength of the sentiments expressed by the writers (Ku et al., 2006).

Extractive summaries are linguistically interesting and can be both informative and concise. Extractive summarizers also require less engineering effort. On the other hand, abstractive summaries tend to have better coverage for a particular level of conciseness, and to be less redundant and more coherent (Carenini et al., 2012). They also can be constructed to target particular discourse goals, such as summarization, comparison or recommendation. Although in theory, it is possible to produce user-targeted extractive summaries, user-specific review summarization has only been explored in the context of abstractive summarization (Carenini et al., 2012).

Current systems for summarizing sentiment-laden text use information about the attributes of the target entity (or entities); the range, mean and median of the ratings of each attribute; relationships between the attributes; and links between ratings/attributes and text elements in the input documents (Blair-Goldensohn et al., 2008). However, there is other information that no summarizer currently takes into account. This includes temporal features (in particular, depending on how old the documents are, products and services evaluated features may change over time) and social features (in particular, social or demographic similarities or relationships between document authors and the reader of the summary). In addition, there is an essential contradiction at the heart of current review summarization systems: the system is authoring the review, but the opinions contained therein are really attributable to one or more human authors, and those attributions are not retained in the review summary. For example, consider the extractive summary generated with STARLET-E (Di Fabbrizio et al., 2013): *"Delicious. Can't wait for my next trip to Buffalo. GREAT WINGS. I have rearranged business trips*

*This work was conducted when in AT&T Labs Research

Proceedings of the 8th International Natural Language Generation Conference, pages 54–63,
Philadelphia, Pennsylvania, 19-21 June 2014.

so that I could stop in and have a helping or two of their wings". We were seated promptly and the staff was courteous.

The summary is generated by selecting sentences from reviews to reflect topics and rating distributions contained in the input review set. Do the two sentences about wings reflect one (repeated) opinion from a single reviewer, or two opinions from two separate reviewers? The ability to attribute subjective statements to known sources can make them more trustworthy; conversely, in the absence of the ability to attribute, a reader may become skeptical or confused about the content of the review summary. We term this summarization issue *opinion holder attribution*.

In this paper we present STARLET-H, a *hybrid* review summarizer that combines the advantages of the abstractive and extractive approaches to summarization and implements a solution to the opinion holder attribution problem. STARLET-H takes as input a set of reviews, each review of which is labeled with aspect ratings and authorship. It generates *hybrid abstractive/extractive* reviews that: 1) are informative (achieve broad coverage of the input opinions); 2) are concise and avoid redundancy; 3) are readable and coherent (of high linguistic quality); 4) can be targeted to the reader; and 5) address the opinion holder attribution problem by directly referring to reviewers authorship when embedding phrases from reviews. We demonstrate through a comparative evaluation of STARLET-H and other review summarizers that hybrid review summarization is preferred over extractive summarization for readability, correctness, completeness (achieving broad coverage of the input opinions) and compactness.

2 Hybrid summarization

Most NLG research has converged around a "consensus architecture" (Reiter, 1994; Rambow and Korelsky, 1992), a pipeline architecture including the following modules: 1) **text planning**, which determines how the presentation content is selected, structured, and ordered; 2) **sentence planning**, which assigns content to sentences, inserts discourse cues to communicate the structure of the presentation, and performs sentence aggregation and optionally referring expression generation; and 3) **surface realization**, which performs lexical selection, resolves syntactic issues such as subject-verb and noun-determiner agreement, and assigns morphological inflection to produce the final grammatical sentence. An abstractive summarizer requires the customization of these three modules. Specifically, the text planner has to select and organize the information contained in the input reviews to reflect the rating distributions over the aspects discussed by the reviewers. The sentence planner must perform aggregation in such a way as to optimize summary length without confusing the reader, and insert discourse cues that reveal the discourse structure underlying the summary. And, finally, the surface realizer must select the proper domain lexemes to express positive and negative opinions.

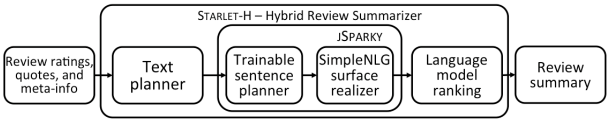

Figure 1: STARLET-H hybrid review summarizer architecture

Figure 1 shows the architecture we adopted for our STARLET-H hybrid review summarizer. We use a generate-and-select approach: the decisions to be made at each stage of the NLG process just outlined are complex, and because they are not truly independent of each other, a generate-and-rank approach may be best (allowing each component to express alternative 'good' choices and choosing the best combination of these choices at the end). Our text planner is responsible for analyzing the input text reviews, extracting per-attribute rating distributions and other meta-data from each review, and synthesizing this information to produce one or more discourse plans. Our sentence planner, JSPARKY – a freely-available toolkit (Stent and Molina, 2009) – can produce several candidate sentence plans and their corresponding surface realizations through SimpleNLG (Gatt and Reiter, 2009). The candidate summaries are ranked by calculating their perplexity with a language model trained over a large number of sentences from additional restaurant reviews collected over the Web.

2.1 Data

STARLET-H uses review data directly, as input to summarization, and indirectly, as training data for statistical models and for lexicons for various stages of the summarization process.

For training data, we used two sets of labeled data: one for the restaurant domain and the other for the hotel domain. Both corpora include manually created sentence-level annotations

that identify: 1) opinion targets – phrases referring to domain-relevant aspects that are the targets of opinions expressed by the reviewer; 2) opinion phrases – phrases expressing an opinion about an entity, and its polarity (positive or negative); and 3) opinion groups – links between opinion phrases and their opinion targets. Additionally, sentences satisfying the properties of quotable sentence mentioned in Section 3 were labeled as "quotable". Table 1 summarizes the overall statistics of the two corpora. The annotated corpora included the following rated aspects: *Atmosphere, Food, Service, Value*, and *Overall* for the Restaurant domain, and *Location, Rooms, Service, Value*, and *Overall* for the Hotel domain[1].

Table 1: Quote-annotated dataset statistics

Dataset	**RQ4000**	**HQ4000**	
Domain	Restaurant	Hotel	Total
Reviews	484	404	888
Sentences	4,007	4,013	8,020
Avg sentences / review	8.28	9.93	9.03

2.2 Text planning

Reviews present highly structured information: each contains an (implicit or explicit) rating of one or more aspects of a target entity, possibly with justification or evidence in the form of examples. The rich information represented in these ratings – either directly expressed in reviews or extracted by an automatic rating prediction model – can be exploited in several ways. Our text planner receives as input a set of text reviews with associated per-aspect ratings, and for each review proceeds through the following analysis steps:

Entity description Extracts basic information to describe the reviewed entity, e.g., the name and location of the business, number of total and recent reviews, review dates and authors, etc.

Aspect distribution categorization Categorizes the rating distribution for each aspect of the reviewed entity as one of four types: 1) positive – most of the ratings are positive; 2) negative – most of the ratings are negative; 3) bimodal – most of the ratings are equally distributed into positive and negative values; 4) uniform – ratings are uniformly distributed across the rating scale.

[1]Some examples from the annotated corpus are available at the following address `http://s286209735.onlinehome.us/starlet/examples`

Quote selection and attribution Classifies each sentence from the reviews using a quote selection model (see Section 3), which assigns to each sentence an aspect, a rating polarity (positive/negative) and a confidence score. The classified sentences are sorted by confidence score and a candidate quote is selected for each aspect of the target entity that is explicitly mentioned in the input reviews. Each quote is stored with the name of the reviewer for correct authorship attribution. Note that when the quote selection module is excluded, the system is an abstractive summarizer, which we call STARLET-A.

Lexical selection Selects a lexicon for each aspect based on its rating polarity and its assigned rating distribution type. Lexicons are extracted from the corpus of annotated opinion phrases described in Di Fabbrizio et al. (2011).

Aspect ordering Assigns an order over aspects using aspect ordering statistics from our training data (see Section 2.4), and generates a discourse plan, using a small set of rhetorical relations organized into summary templates (see below).

2.3 Sentence planning

The STARLET-H sentence planner relies on rhetorical structure theory (RST) (Mann and Thompson, 1989). RST is a linguistic framework that describes the structure of natural language text in terms of the *rhetorical* relationships organizing textual units. Through a manual inspection of our training data, we identified a subset of six RST relations that are relevant to review summarization: *concession, contrast, example, justify, list*, and *summary*. We further identified four basic RST-based summary templates, one for each per-aspect rating distribution: mostly positive, mostly negative, uniform across all ratings, and bimodal (e.g., both positive and negative). These summary templates are composed by the text planner to build summary discourse plans. The JSPARKY sentence planner then converts input discourse plans into sentence plans, performing sentence ordering, sentence aggregation, cross-sentence reference resolution, sentence tense and mode (passive or active), discourse cue insertion, and the selection of some lexical forms from FrameNet (Baker et al., 1998) relations.

Figure 2 illustrates a typical RST template representing a positive review summary and corresponding text output generated by JSPARKY. For each aspect of the considered domain, the sentence plan strategy covers a variety of opinion distribu-

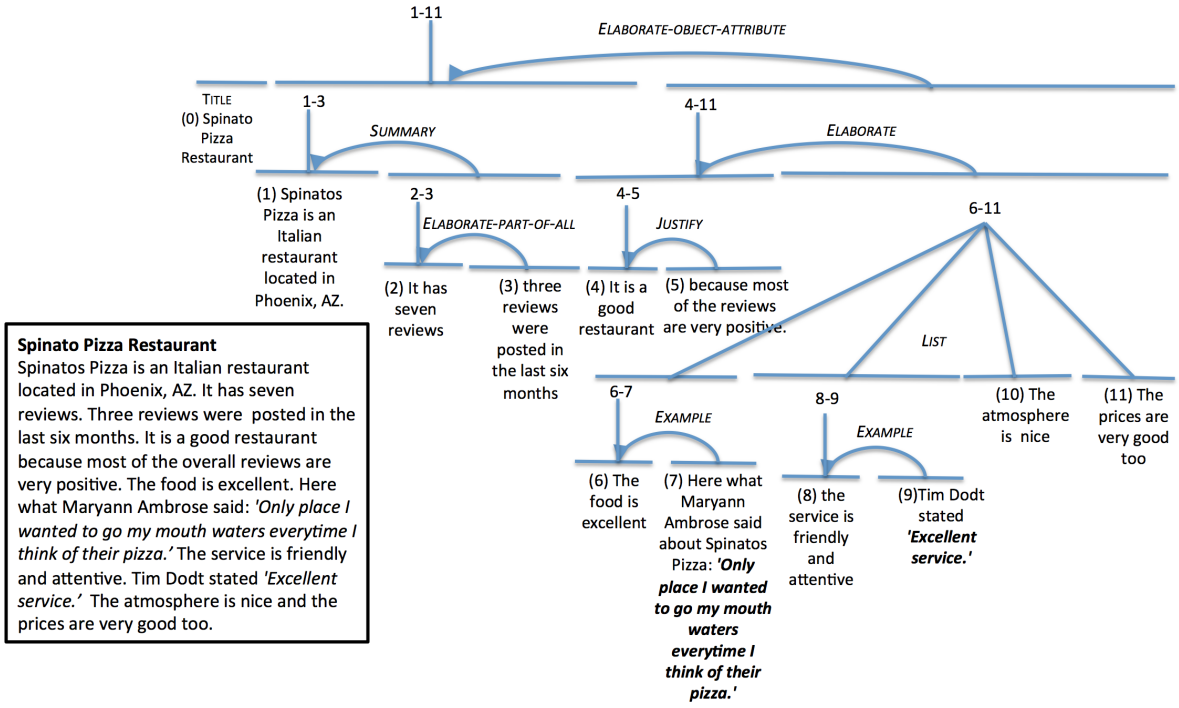

Figure 2: Example of RST structure generated by the text planner for mostly positive restaurant reviews

tion conditions (e.g., positive, negative, bimodal, and uniform), and provides alternative RST structures when the default relation is missing due to lack of data (e.g., missing quotes for a specific aspect, missing information about review distribution over time, missing type of cuisine, and so on). The sentence template can also manage lexical variations by generating multiple options to qualify a specific pair of aspect and opinion polarity. For instance, in case of very positive reviews about restaurant atmosphere, it can provide few alternative adjective phrases (e.g., great, wonderful, very warm, terrific, etc.) that can be used to produce more summary candidates (over-generate) during the final surface realization stage.

2.4 Ordering aspects and polarities

The discourse structure of a typical review consists of a summary opinion, followed by a sequence of per-aspect ratings with supporting information (e.g., evidence, justification, examples, and concessions). The preferred sequence of aspects to present in a summary depends on the specific review domain, the overall polarity of the reviews, and how opinion polarity is distributed across the reviewed aspects. Looking at our training data, we observed that when the review is overall positive, positively-rated aspects are typically discussed at the beginning, while negatively-rated aspects tend to gather toward the end. The opposite order seems predominant in the case of negative reviews. When opinions are mixed, aspect ordering strategies are unclear. To most accurately model aspect ordering, we trained weighted finite state transducers for the restaurant and hotel domains using our training data. Weighted finite state transducers (WFSTs) are an elegant approach to search large feature spaces and find optimal paths by using well-defined algebraic operations (Mohri et al., 1996). To find the optimal ordering of rated aspects in a domain, the text planner creates a WFST with all the possible permutations of the input sequence of aspects, and composes it with a larger WFST trained from bigram sequences of aspects extracted from the relevant domain-specific review corpus. The best path sequence is then derived from the composed WFST by applying the Viterbi decoding algorithm. For instance, the sequence of aspects and polarities represented by the string: `value-n service-p overall-n food-n atmosphere-n`[2] is first permuted in all the different possible sequences and then converted into a WFST. Then the permutation network is fully composed with the larger, corpus-trained WFST. The best path is extracted by dynamic programming, producing the optimal sequence `service-p value-n overall-n atmosphere-n food-n`.

[2]We postfix the aspect label with a '-p' for positive and with '-n' for negative opinion

2.5 Lexical choice

It can be hard to choose the best opinion words, especially when the summary must convey the different nuances between "good" and "great" or "bad" and "terrible" for a particular aspect in a particular domain. For our summarization task, we adopted a simple approach. From our annotated corpora, we mined both positive and negative opinion phrases with their associated aspects and rating polarities. We sorted the opinion phrases by frequency and then manually selected from the most likely phrases adjective phrases that may correctly express per-aspect polarities. We then split positive and negative phrases into two levels of polarity (i.e., strongly positive, weakly positive, weakly negative, strongly negative) and use the number of star ratings to select the right polarity during content planning. For bimodal and uniform polarity distributions, we manually defined a customized set of terms. Sample lexical terms are reported in Table 2.

3 Quote selection modeling

There are several techniques to extract salient phrases from text, often related to summarization problems, but there is a relatively little work on extracting *quotable sentences* from text (Sarmento and Nunes, 2009; De La Clergerie et al., 2009) and none, to our knowledge, on extracting quotes from sentiment-latent text. So, what does make a phrase quotable? What is a proper quote definition that applies to review summarization? We define a *sentiment-laden quotable phrase* as a text fragment with the following characteristics: **attributable** – clearly ascribable to the author; **compact and simple** – it is typically a relatively short phrase (between two and twenty words) which contains a statement with a simple syntactic structure and independent clauses; **self-contained** its meaning is clear and self-contained, e.g., it does not include pronominal references to entities outside its scope; **on-topic** – it refers to opinion targets (i.e., aspects) in a specific domain; **sentiment-laden** – it has one or two opinion targets and an unambiguous overall polarity. Example quotable phrases are presented in Table 3.

To automatically detect quotes from reviews, we adopted a supervised machine learning approach based on manually labeled data. The classification task consists of classifying both aspects and polarity for the most frequent aspects defined for each domain. Quotes for the aspect `food`, for instance, are split into positive and negative classification labels: `food-p` and `food-n`, respectively. We identify quotable phrases and associate them with aspects and rating polarities all in one step, but multi-step approaches could also be used (e.g., a configuration with binary classification to detect quotable sentences followed by another classification model for aspect and polarity detection).

Table 3: Example of quotes from restaurant and hotel domains

'Everyone goes out of their way to make sure you are happy with their service and food.'
'The stuffed mushrooms are the best I've ever had as was the lasagna.'
'Service is friendly and attentive even during the morning rush.'
'I've never slept so well away from home loved the comfortable beds.'
'The price is high for substandard mattresses when I pay this much for a room.'

3.1 Training quote selection models

We used the following features for automatic quote selection: **ngrams** – unigrams, bigrams, and trigrams from the input phrases with frequency higher than three; **binned number of words** – we assumed a maximum length of twenty words per sentence and created six bins, five of them uniformly distributed from one to twenty, and the sixth including all the sentences of length greater than twenty words; **POS** – unigrams, bigrams, and trigrams for part of speech tags; **chunks** – unigrams, bigrams, and trigrams for shallow parsed syntactic chunks; **opinion phrases** – a binary feature to keep track of the presence of positive and negative opinion phrases as defined in our annotated review corpora. In our annotated data only the most popular aspects are well represented. For instance, `food-p` and `overall-p` are the most popular positive aspects among the quotable sentences for the restaurant domain, while quotes on `atmosphere-n` and `value-n` are scarce. The distribution is even further skewed for the hotel domain; there are plenty of quotes for `overall-p` and `service-p` and only 13 samples (0.43%) for `location-n`. To compensate for the broad variation in the sample population, we used *stratified sampling* methods to divide the data into more balanced testing and training data We generated 10-fold stratified training/test sets. We experimented with three machine learning algorithms: MaxEnt, SVMs with linear kernels, and SVMs with polynomial kernels. The MaxEnt learning algorithm produced statistically better classification results than the other algorithms when used with uni-

Table 2: Summarizer lexicon for most frequent adjective phrases by aspect and polarity

Domain	Restaurant				Hotel				
Aspect	positive	very positive	negative	very negative	Aspect	positive	very positive	negative	very negative
atmosphere	nice, good, friendly, comfortable	great, wonderful, very warm, terrific	ordinary, depressing	really bad	location	good, nice, pleasant	amazing, awesome, excellent, great	bad, noisy, gloomy	very bad, very bleak, very gloomy
food	good, delicious, pleasant, nice, hearty, enjoyable	great, excellent, very good, to die for, incredible	very basic, unoriginal, uninteresting, unacceptable, substandard, poor	mediocre, terrible, horrible, absolutely horrible	rooms	comfortable, decent, clean, good	amazing, awesome, gorgeous	average, basic, subpar	terrible, very limited, very average
overall	good, quite enjoyable, lovely	wonderful, terrific, very nice	bad, unremarkable, not so good	absolutely terrible, horrible, pretty bad	overall	great, nice, welcoming	excellent, superb, perfect	average, nothing great, noisy	quite bad, awful, horrible
service	attentive, friendly, pleasant, courteous	very attentive, great, excellent, very friendly	inattentive, poor, not friendly, bad	extremely poor, horrible, so lousy, awful	service	friendly, great, nice, helpful, good	very friendly, great, excellent, very nice	average, basic, not that great	very bad, dreadful
value	reasonable, fair, good value	very reasonable, great	not that good, not worthy	terrible, outrageous	value	great, nice, good, decent	very good, wonderful, perfectly good	not good	not very good

gram features. This confirmed a general trend we have previously observed in other text classification experiments: with relatively small and noisy datasets, unigram features provide better discriminative power than sparse bigrams or trigrams, and MaxEnt methods are more robust when dealing with noisy data.

3.2 Quote selection results

Table 4 reports precision, recall and F-measures averaged across 10-fold cross-validated test sets with relative standard deviation. The label `nq` identifies non-quotable sentences, while the other labels refer to the domain-specific aspects and their polarities. For the quote selection task, precision is the most important metric: missing some potential candidates is less important than incorrectly identifying the polarity of a quote or substituting one aspect with another. The text planner in STARLET-H further prunes the quotable phrases by considering only the quote candidates with the highest scores.

4 Evaluation

Evaluating an abstractive review summarizer involves measuring how accurately the opinion content present in the reviews is reflected in the summary and how understandable the generated content is to the reader. Traditional multi-document summarization evaluation techniques utilize both qualitative and quantitative metrics. The former require human subjects to rate different evaluative characteristics on a Likert-like scale, while the latter relies on automatic metrics such as ROUGE (Lin, 2004), which is based on the common number of n-grams between a peer, and one or several gold-standard reference summaries.

Table 4: Quote, aspect, and polarity classification performances for the restaurant domain

	Precision	Recall	F-measure
atmosphere-n	0.233	0.080	0.115
atmosphere-p	0.589	0.409	0.475
food-n	0.634	0.409	0.491
food-p	0.592	0.634	0.612
nq	0.672	0.822	0.740
overall-n	0.545	0.275	0.343
overall-p	0.555	0.491	0.518
service-n	0.699	0.393	0.498
service-p	0.716	0.563	0.626
value-n	0.100	0.033	0.050
value-p	0.437	0.225	0.286
Hotel	**Precision**	**Recall**	**F-measure**
location-n	-	-	-
location-p	0.572	0.410	0.465
nq	0.678	0.836	0.748
overall-n	0.517	0.233	0.305
overall-p	0.590	0.492	0.536
rooms-n	0.628	0.330	0.403
rooms-p	0.667	0.573	0.612
service-n	0.517	0.163	0.240
service-p	0.605	0.500	0.543
value-n	-	-	-
value-p	0.743	0.300	0.401

4.1 Evaluation materials

To evaluate our abstractive summarizer, we used a qualitative metric approach and compared four review summarizers: 1) the open source MEAD system, designed for extractive summarization of general text (Radev et al., 2004); 2) STARLET-E, an extractive summarizer based on KL-divergence and language modeling features that is described in Di Fabbrizio et al. (2011); 3) STARLET-A, the abstractive summarizer presented in this paper, *without* the quote selection module; and 4) the hybrid summarizer STARLET-H.

We used the Amazon Mechanical Turk[3] crowd-

[3] `http://www.mturk.com`

sourcing system to post subjective evaluation tasks, or HITs, for 20 restaurant summaries. Each HIT consists of a set of ten randomly ordered reviews for one restaurant, and four randomly ordered summaries of reviews for that restaurant, each one accompanied by a set of evaluation widgets for the different evaluation metrics described below. To minimize reading order bias, both reviews and summaries were shuffled each time a task was presented.

4.2 Evaluation metrics

We chose to carry out a qualitative evaluation in the first instance as n-gram metrics, such as ROUGE, are not necessarily appropriate for assessing abstractive summaries. We asked each participant to evaluate each summary by rating (using a Likert scale with the following rating values: 1) Not at all; 2) Not very; 3) Somewhat; 4) Very; 5) Absolutely) the following four summary criteria: **readability** – a summary is readable if it is easy to read and understand; **correctness** – a summary is correct if it expresses the opinions in the reviews; **completeness** – a summary is complete if it captures the whole range of opinions in the reviews; **compactness** – a summary is compact if it does not repeat information.

4.3 Evaluation procedure

We requested five evaluators for each HIT. To increase the chances of getting accurate evaluations, we required evaluators to be located in the USA and have an approval rate of 90% or higher (i.e., have a history of 90% or more approved HITs). Manual examinations of the evaluation responses did not show evidence of tampered data, but statistical analysis showed unusually widely spread rating ranges. We noticed that most evaluators only evaluated one or two HITs; this may imply that they tried a few HITs and then decided not to continue because they found the task too long or the instructions unclear. We then re-opened the evaluation and directly contacted three additional evaluators, explaining in detail the instructions and the evaluation scales. For consistency, we asked these evaluators to complete the evaluation for all HITs. In our analysis, we only included the five evaluators (two from the first round of evaluation, and three from the second) who completed all HITs. For each evaluation metric, the five workers evaluated each of the 20 summaries, for a total of 100 ratings. Table 5 shows an example output of the four summarization methods for a single set of restaurant review documents.

Table 5: Example of MEAD-based, extractive, abstractive and hybrid summaries from the restaurant domain

MEAD Summary a truly fun resturant everyone who like spicy food should try the rattoes and for a mixed drink the worm burner really good food and a fun place to meet your friends. We were attracted by the great big frog on the exterior of the building and the fun RAZZOO S logo during a trip to the mall. it was great the waitress was excellent very prompt and courteous and friendly to all a real complement to razzoo 's way of service her name was Tabitha. The best spicy food restaurant with great server and fast service.
Extractive summary Eat there every chance i get. We ve been going here for years. Their crawfish etoufee is the BEST. And such an awesome value for under 10. Excellent as always. Some of the best food in the area. I use to work at Razzoo s. It was hard to leave. The people are great and so is the food. I still go in there and miss it more everytime. I Love Loney. It was great. Our server was great and very observant. Try the Chicken Tchoupitoulas.
Abstractive summary Razzoo's Cajun Cafe in Concord, NC is an American restaurant. It has nine reviews. It had three very recent reviews. It is an awesome, American restaurant. It has many very positive reviews. It has an excellent atmosphere and and has always exceptional service.
Hybrid summary Razzoo's Cajun Cafe in Concord, NC is an American restaurant. It has nine reviews. It had three very recent reviews. It is an awesome, American restaurant. It has many very positive reviews. First it has a great price. Angela Haithcock says ''And such an awesome value for under 10''. Second it has always exceptional service and for instance Danny Benson says ''it was great the waitress was excellent very prompt and courteous and friendly to all a real complement to razzoo's way of service her name was Tabitha''. Third it has an excellent atmosphere. Last it has amazing food. Scott Kern says ''Some of the best food in the area''.

4.4 Evaluation results and discussion

The evaluation results are presented in Table 6. Each evaluation metric is considered separately. Average values for STARLET-E, STARLET-A and STARLET-H are better than for MEAD across the board, suggesting a preference for summaries of sentiment-laden text that take opinion into account. To validate this hypothesis, we first computed the non-parametric Kruskal-Wallis statistic for each evaluation metric, using a chi-square test to establish significance. The results were not significant for any of the metrics.

However, when we conducted pairwise Wilcoxon signed-rank tests considering two summarization methods at a time, we found some significant differences ($p < 0.05$). As predicted,

Table 6: Qualitative evaluation results

	MEAD	Starlet-E	Starlet-A	Starlet-H
Readability	2.95	3.17	3.64	3.74
Completeness	2.88	3.29	3.290	3.58
Compactness	3.07	3.35	3.80	3.58
Correctness	3.26	3.48	3.59	3.72

MEAD perform substantially worse than both STARLET-A and STARLET-H on readability, correctness, completeness, and compactness. STARLET-A and STARLET-H are also preferred over STARLET-E for readability. While STARLET-A is preferred over STARLET-E for compactness (the average length of the abstractive reviews was 45.05 words, and of the extractive,102.30), STARLET-H is preferred over STARLET-E for correctness, since the former better captures the reviewers opinions by quoting them in the appropriate context. STARLET-A and STARLET-H achieve virtually indistinguishable performance on all evaluation metrics. Our evaluation results accord with those of Carenini et al. (2012); their abstractive summarizer had superior performance in terms of content precision and accuracy when compared to summaries generated by an extractive summarizer. Carenini et al. (2012) also found that the differences between extractive and abstractive approaches are even more significant in the case of controversial content, where the abstractive system is able to more effectively convey the full range of opinions.

5 Related work

Ganesan et al. (2010) propose a method to extract salient sentence fragments that are both highly frequent and syntactically well-formed by using a graph-based data structure to eliminate redundancies. However, this approach assumes that the input sentences are already selected in terms of aspect and with highly redundant opinion content. Also, the generated summaries are very short and cannot be compared to a full-length output of a typical multi-document summarizer (e.g., 100-200 words). A similar approach is described in Ganesan et al. (2012), where very short phrases (from two to five words) are collated together to generate what the authors call *ultra-concise* summaries.

The most complete contribution to evaluative text summarization is described in Carenini et al. (2012) and it closely relates to this work. Carenini et al. (2012) compare an extractive summarization system, MEAD* – a modified version of the open source summarization system MEAD (Radev et al., 2004) – with SEA, an abstractive summarization system, demonstrating that both systems perform equally well. The SEA approach, although better than traditional MEAD, has a few drawbacks. Firstly, the sentence selection mechanism only considers the most frequently discussed aspects, leaving the decision about where to stop the selection process to the maximum summary length parameter. This could leave out interesting opinions that do not appear with sufficient frequency in the source documents. Ideally, all opinions should be represented in the summary according to the overall distribution of the input reviews. Secondly, Carenini et al. (2012) use the absolute value of the sum of positive and negative contributions to determine the relevance of a sentence in terms of opinion content. This flattens the aspect distributions since sentences with very negative or very positive polarity or with numerous opinions, but with moderate polarity strengths, will get the same score, regardless. Finally, it does not address the opinion holder attribution problem leaving the source of opinion undefined. In contrast, STARLET-H follows reviews aspect rating distributions both to select quotable sentences and to summarize relevant aspects. Moreover, it explicitly mentions the opinion source in the embedded quoted sentences.

6 Conclusions

In this paper, we present a hybrid summarizer for sentiment-laden text that combines an overall abstractive summarization method with an extractive summarization-based quote selection method. This summarizer can provide the readability and correctness of abstractive summarization, while addressing the opinion holder attribution problem that can lead readers to become confused or misled about who is making claims that they read in review summaries. We plan a more extensive evaluation of STARLET-H. Another potential area of future research concerns the ability to personalize summaries to the user's needs. For instance, the text planner can adapt its communicative goals based on polarity orientation – a user can be more interested in exploring in detail negative reviews – or it can focus more on specific (user-tailored) aspects and change the order of the presentation accordingly. Finally, it could be interesting to customize the summarizer to provide an overview of what is available in a specific geographic neighborhood and compare and contrast the options.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The Berkeley FrameNet Project. In *Proceedings of the 17th International Conference on Computational Linguistics - Volume 1*, COLING '98, pages 86–90, Stroudsburg, PA, USA, 1998. Association for Computational Linguistics. doi: 10.3115/980451.980860. URL `http://dx.doi.org/10.3115/980451.980860`.

Sasha Blair-Goldensohn, Kerry Hannan, Ryan McDonald, Tyler Neylon, George Reis, and Jeff Reynar. Building a Sentiment Summarizer for Local Service Reviews. In *NLP in the Information Explosion Era*, 2008.

Giuseppe Carenini, Jackie Chi Kit Cheung, and Adam Pauls. Multi-Document Summarization of Evaluative Text. *Computational Intelligence*, 2012.

Éric De La Clergerie, Benoît Sagot, Rosa Stern, Pascal Denis, Gaëlle Recourcé, and Victor Mignot. Extracting and Visualizing Quotations from News Wires. In *Language and Technology Conference*, Poznan, Pologne, 2009. Projet Scribo (pôle de compétitivité System@tic).

Giuseppe Di Fabbrizio, Ahmet Aker, and Robert Gaizauskas. STARLET: Multi-document Summarization of Service and Product Reviews with Balanced Rating Distributions. In *Proceedings of the 2011 IEEE International Conference on Data Mining (ICDM) Workshop on Sentiment Elicitation from Natural Text for Information Retrieval and Extraction (SENTIRE)*, Vancouver, Canada, december 2011.

Giuseppe Di Fabbrizio, Ahmet Aker, and Robert Gaizauskas. Summarizing On-line Product and Service Reviews Using Aspect RatingDistributions and Language Modeling. *Intelligent Systems, IEEE*, 28(3):28–37, May 2013. ISSN 1541-1672. doi: 10.1109/MIS.2013.36.

Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. Opinosis: A Graph-Based Approach to Abstractive Summarization of Highly Redundant Opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 340–348, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

Kavita Ganesan, ChengXiang Zhai, and Evelyne Viegas. 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, pages 869–878, New York, NY, USA, 2012. ACM.

Albert Gatt and Ehud Reiter. SimpleNLG: A Realisation Engine for Practical Applications. In *Proceedings of the 12th European Workshop on Natural Language Generation*, ENLG '09, pages 90–93, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz. Multi-document Summarization by Sentence Extraction. In *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic summarization - Volume 4*, pages 40–48, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.

Lun-Wei Ku, Yu-Ting Liang, and Hsin-Hsi Chen. Opinion Extraction, Summarization and Tracking in News and BlogCorpora. In *Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs*, 2006.

Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of summaries. In *Proc. ACL workshop on Text Summarization Branches Out*, page 10, 2004.

William C. Mann and Sandra A. Thompson. Rhetorical Structure Theory: A Theory of Text Organization. In Livia Polanyi, editor, *The Structure of Discourse*. Ablex, Norwood, NJ, 1989.

Mehryar Mohri, Fernando Pereira, and Michael Riley. Weighted Automata in Text and Speech Processing. In *ECAI-96 Workshop*, pages 46–50. John Wiley and Sons, 1996.

Dragomir Radev, Timothy Allison, Sasha Blair-Goldensohn, John Blitzer, Arda Çelebi, Stanko Dimitrov, Elliott Drabek, Ali Hakim, Wai Lam, Danyu Liu, Jahna Otterbacher, Hong Qi, Horacio Saggion, Simone Teufel, Michael Topper, Adam Winkel, and Zhu Zhang. MEAD – A Platform for Multidocument Multilingual Text Summarization. In *Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal, May 2004.

Dragomir R. Radev and Kathleen R. McKeown. Generating natural language summaries from multiple on-line sources. *Computational Linguistiscs*, 24(3):470–500, September 1998. ISSN 0891-2017.

Owen Rambow and Tanya Korelsky. Applied Text Generation. In *Proceedings of the Third Conference on Applied Natural Language Processing*, pages 40–47, Trento, Italy, 1992. Association for Computational Linguistics. 31 March - 3 April.

Ehud Reiter. Has a Consensus NL Generation Architecture Appeared, and is it Psychologically Plausible? In David McDonald and Marie Meteer, editors, *Proceedings of the 7th. International Workshop on Natural Language generation (INLGW '94)*, pages 163–170, Kennebunkport, Maine, 1994.

Luis Sarmento and Sérgio Nunes. Automatic Extraction of Quotes and Topics from News Feeds. In *4th Doctoral Symposium on Informatics Engineering (DSIE09)*, 2009.

Amanda Stent and Martin Molina. Evaluating Automatic Extraction of Rules for Sentence Plan Construction. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '09, pages 290–297, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.