2011 11th IEEE International Conference on Data Mining Workshops

STARLET: Multi-document Summarization of Service and Product Reviews with Balanced Rating Distributions

Giuseppe Di Fabbrizio*†, Ahmet Aker*, Robert Gaizauskas*
* *Department of Computer Science*
University of Sheffield, Sheffield, S1 4DP - UK
Email: {G.DiFabbrizio,A.Aker,R.Gaizauskas}@sheffield.ac.uk
† *AT&T Labs - Research*
Florham Park, NJ 07932 - USA
Email: pino@research.att.com

***Abstract*—Reviews about products and services are abundantly available online. However, selecting information relevant to a potential buyer involves a significant amount of time reading user's reviews and weeding out comments unrelated to the important aspects of the reviewed entity. In this work, we present STARLET, a novel approach to multi-document summarization for evaluative text that considers the rating distribution as summarization feature to consistently preserve the overall opinion distribution expressed in the original reviews. We demonstrate how this method improves traditional summarization techniques and leads to more readable summaries.**

***Keywords*-Summarization; evaluative text; A* search; multi-ratings prediction**

I. INTRODUCTION

With the broad availability of always-connected portable devices such as mobiles, tablets, and eReaders, condensing information for displaying in a relatively small screen has become a necessity for the exceedingly demanding population of users *on-the-go*. Retail industry and service providers are recognizing that there is a growing crowd of potential customers who are relying on their devices to learn about products and services, discover other user's experiences, and, ultimately, make a decision about spending their money or not [1], [2].

Although reviews about products and services are abundantly available online, selecting information relevant to a potential buyer involves a significant amount of time reading reviews and weeding out comments unrelated to the important aspects of the reviewed entity. This is particularly true for mobile users where additional constraints about geographic location together with limited screen size may affect consumer's purchase behavior. In order to summarize reviews on mobile devices, several steps are required, each one involving different and often poorly integrated technology.

For instance, a first step in such a process would involve the use of *local mobile search* techniques [3] to find services or products available nearby. In this case, the search engine will have to re-rank output results by using geographic information about the current user's location [4] - or an explicitly requested location - and, optionally, re-score the final results based on previous search history captured in the user profile. Next, the user will start evaluating the search results by exploring reviews and ratings posted online by other users. In this case, *opinion mining* and *sentiment analysis* methods can be applied to extract the target of the opinions expressed in the reviews and the relative polarity (e.g., positive, negative, or neutral) [5], [6], [7]. Lastly, the user will be engaged in a complex task to process all the facts, opinions, and ratings read in the previous step and subsequently interpret, compare, contrast, and, finally, *summarize* the needed information.

While the first two steps have been largely explored, summarization of evaluative text (e.g., documents containing opinion or sentiment-laden text), is fairly new and may be substantially different from the traditional summarization task. In fact, most summarization techniques focus on distilling factual information by identifying the documents main topics, removing redundancies, and coherently ordering the extracted phrases or sentences. Most of the contributions have been developed using corpora with well-formed documents from domains such as news articles [8], [9], medical literature [10], biographies [11], technical articles [12], and blogs [13]. As observed in Ku et al. [14], traditional summarization tends to identify and discard redundancies, while in sentiment-laden text, similar opinions mentioned multiple times across documents are crucial indicators of the overall strength of the sentiments expressed by the writers.

More specifically, sentiment-laden documents like product and service reviews are usually either about a single *entity*, e.g., consumers' goods such as digital cameras, DVD players, books; or related to user's experiences of a *service* like lodging in an hotel or dining in a restaurant. Typically an entity has several ratable features or *aspects* which may be targeted by reviewers with their positive or negative

978-0-7695-4409-0/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDMW.2011.158

opinions. In this sense, each review can be viewed as a set of aspects with associated opinions. Ratings define the strength and the polarity of the opinions and typically range over integer values often visualized with star symbols. When summarizing reviews, it is fundamental to identify the opinion information expressed across the reviews and their polarity distribution, so that the sentences selected by a summarizer could be representative of the overall sentiment distribution. In addition to the traditional tasks, a multi-document opinion-oriented summarizer requires an information extraction stage to identify topics and polarity described in the documents.

In the service domain case, many web sites[1] allow reviewers to directly rate pre-defined aspects. E.g., for restaurants typical aspects are *atmosphere, food, value, service*, and *overall* with ratings ranging from *poor* (one star) to *excellent* (five stars). These rated aspects quantify opinions and polarities expressed in each review by the reviewers, and although there might be inconsistencies, it is safe to assume that the text document associated with the ratings carries the same sentiment contributions quantified by the number of stars. By the same token, aggregating the ratings of the single reviews over the aspects can yield a fair summary of the overall sentiments expressed by the reviewers of the specific service or entity reviewed. Based on these considerations, we can assume that a summary should convey the same distribution of ratings over aspects obtained by combining the rating contribution of each review, so that each opinion contribution, even if controversial, should be represented into the final summary.

This work proposes STARLET, a novel approach to summarization of evaluative text that utilizes aspects and ratings described in the reviews as features for the summarization process. In the restaurant domain, which we investigate as an example domain for service reviews, STARLET uses *atmosphere, food, value, service*, and *overall* aspects to score each sentence in the input documents. For each aspect, STARLET computes a rating indicating how much the current sentence has contributed to that aspect. For this STARLET uses a maximum entropy rating model. The predicted aspect ratings are used in a summarization model to (1) compute a score for each sentence and (2) to derive a summary score. The model is a linear weighted model with aspects as features and associated weights learned using A* search and discriminative training [15].

The rest of this paper is structured as follows: Section II reports current contributions to summarization of evaluative text. Section III describes the summarization model, while Section IV outlines the weight learning procedure. Section V illustrates the aspect rating prediction model, which is followed by a description of the experimental setup in Section VI. We presents the results in Section VII and conclude with Section VIII.

[1] we8there.com, tripadvisor.com, citysearch.com

II. BACKGROUND

From a high level point of view, approaches to multi-document summarization can be divided into *extractive* or *abstractive*. Extractive summarization assumes that fragments (phrases or sentences) extracted from the original documents can be used to assemble a *coherent* shorter version of the original text without substantially changing the information conveyed by the source. Abstractive summarization generates new documents by analyzing the semantic content of the original documents and using natural language generation techniques. While both types of summarization have been extensively studied for factual and edited text documents, there are few contributions extending these approaches to evaluative text summarization. Most of the contributions focus on sentiment analysis and information extraction neglecting how to adapt content selection when sentiment-laden sentences are present.

Features or aspects extraction, for instance, has been explored from many angles: topic models [16], [17], NLP-based information extraction [18], [19], semi-supervised and supervised machine learning techniques [20]. Similarly, for polarity strength prediction (NLP-based [21]) and multi-aspect multi-rating prediction (regression and classification models [22], [23], [24]) are used. Most of the evaluative text summarization methods try to organize the sentiment-laden sentences according to aspect and polarity. In Blair et al. [25], sentences are qualitatively aggregated by aspect and "star ratings" based on a manually-defined strategy; Hu and Liu [7] simply list the sentences by aspects and polarity; in [18], [26] aspect and polarities are graphically organized and visualized. None of these approaches considers text summaries in terms of rating distributions nor introduces metrics to quantitatively evaluate the quality of the summary. To our knowledge, the most complete contribution to evaluative text summarization is described in Carenini et al. [27] and it closely relates to this work.

In [27], Carenini et al. compare an extractive summarization system, MEAD* – a modified version of the open source summarization system MEAD [28] – with SEA, an abstractive summarization system, demonstrating that both systems perform equally well. We only consider extractive summarization.

As noticed in [27], none of the sentence extraction and ranking techniques used in MEAD take into account the sentiment-laden content present in the source documents. Carenini et al. observed that the most informative sentences are the one with the highest number of *crude features* (CF), where CF are defined as the rated aspects of the entity evaluated in the reviews. For each sentence s_k, this score is quantified by the following summation:

$$CF_{sum}(s_k) = \sum_{ps_i \in eval(s_k)} |ps_i| \quad (1)$$

where $eval(s_k)$ is the set of crude features with sentiment-laden content in the sentence, and $|ps_i|$ is the absolute value of the polarity of the crude features referred to in the sentence. For instance, crude features for a digital camera may include *battery life*, *zoom*, *picture quality*, etc. If sentence k mentions positively (+3) the zoom and negatively (-2) the picture quality, the crude feature score would be: $CF_{sum}(s_k) = |+3| + |-2| = 5$. CF features are used in a modified version of MEAD, called MEAD*, to rescore sentences in the final stage of content selection. For this purpose, sentences are inserted and sorted by score in CF 'buckets'. From the CF 'buckets' with more items, MEAD* extracts the sentence with the highest score and removes it from the pool to avoid redundancy. In case two sentences have the same score, the centroid-based feature from MEAD is used as a 'tie-breaker'. The selected sentences are finally ordered according to a pre-defined taxonomy of features where a depth-first traversal of the hierarchy makes sure that aspects are ordered from general to specific.

This 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, using the absolute value when calculating $CF_{sum}(s_k)$ flattens the opinion distribution since sentences with very negative or very positive polarity or with numerous opinions, but with moderate polarity strengths, would get the same score, regardless. Finally, how to use the summarization features is established *a priori* based on expert knowledge and prior work in this area rather than weighting these features from data using automatic quality metrics. In the next few sections, we will address some of these limitations.

III. Summarization Model

Our summarization model s is an adaptation of the one described in Aker et al. [15]. With this model, each possible text summary can be scored as a weighted sum of features according to the equation below:

$$s(\mathbf{y}|\mathbf{x}) = \Phi(\mathbf{y}|\mathbf{x}) \cdot \mathbf{\Lambda} \quad (2)$$

where $\mathbf{x}$ is a vector of indexes representing the N sentences in the document set to summarize, $\mathbf{y} \subseteq \{1, \ldots, N\}$ is the set of indexes selected for the summary of length $|\mathbf{y}| = M$, $\mathbf{\Lambda} = \{\lambda_1, \ldots, \lambda_F\}$ is the weight vector of parameters for the F features that optimizes the summary evaluation metrics, and $\Phi(\cdot|\cdot)$ is a function that returns a set of features for each candidate summary. We assume that all our features can be determined independently[2] leading us to the following score function which is now only dependent on a linear combination of feature functions:

$$s(\mathbf{y}|\mathbf{x}) = \sum_{i \in \mathbf{y}} \phi(x_i)\lambda_i \quad (3)$$

We can finally formulate the summarization task as search problem, where the optimal summary is the one that maximizes the following $\arg\max$ decision rule:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} s(\mathbf{y}|\mathbf{x}) \quad (4)$$

The challenge is to find the scoring parameters $\mathbf{\Lambda}$ that produce a summary $\hat{\mathbf{y}}$ that best approximates an ideal summary $\mathbf{y}$, when compared to a gold standard reference summary using an automatic evaluation metric.

IV. Feature Weight Learning

An extractive multi-document summary can be created by traversing a directed acyclic graph where each node i represents a particular summary of length $l(i)$ composed by a set of sentences S_i, and a set of edges (i, j). Traversing the edge (i, j) incrementally adds a sentence from the set of available sentences to the previous sentence set S_i. Figure 1 shows a graphical representation of the process of selecting sentences. Each node in the graph can be evaluated by a scoring function which quantifies *how good* the node is when compared to a target node. In order to find the best scoring summary with a specific word length W, it would be necessary to search an exponentially large space with complexity $\mathcal{O}(S^{L(W)})$ where S is the total number of sentences to search and $L(W)$ is the number of sentences that best matches the required summary word length W.

The A* search algorithm can be used to efficiently traverse the graph and accurately find the optimal path. It applies a best-first strategy to traverse the graph from the initial node (summary of length zero) to the final node (summary of length W), and uses a heuristic function to determine the order of the nodes to explore first. The search algorithm is guaranteed to converge to the optimal solution if the heuristic function is *monotonic* or follows the *admissible heuristic* requirements. That is, the estimating path cost function from the current node to the goal never overestimates the actual cost. We used the "final aggregated heuristic" function described in [15] that satisfies the admissible heuristic constraints. The input to the heuristic is the set of sentences sorted according to their scores.[3] The heuristic

[2]This assumption does not take into consideration global features such as redundancy or coherence, but the approach can be easily extended to remove this limitation.

[3]Sentence scores are computed by the weighted linear combination of their features.

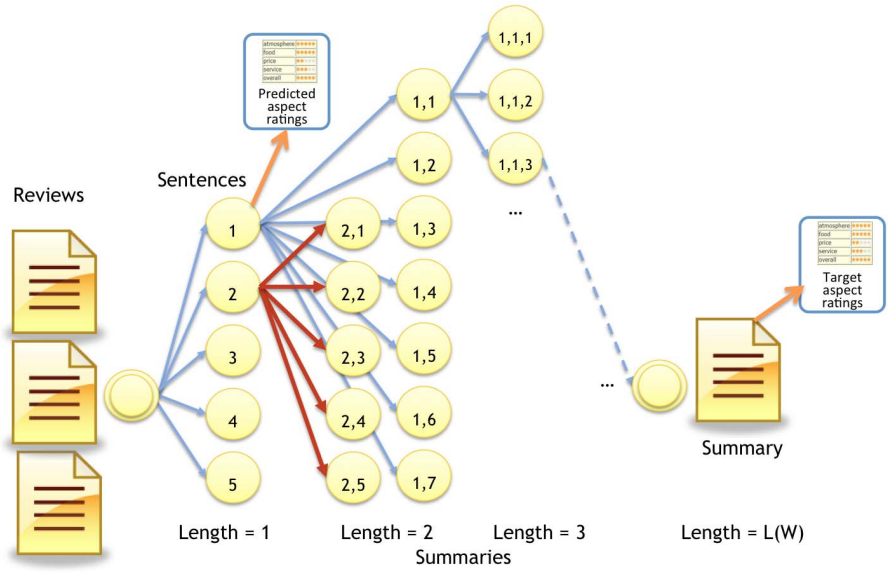

Figure 1. Creating an extractive summary by traversing a directed acyclic graph

first adds the highest scoring sentence into the summary. After adding a sentence, the summary length is updated. If the length limit of the summary is not violated then the next highest scoring sentence is added. When the next high scoring sentence is too long to be added to the summary the heuristic skips this and continues with the next one until it finds the best scoring sentence that does fit into the summary. We generated summaries with a 100-word limit based on the average length of summaries in the reference data.

A. Discriminative Training

After a summary is generated by A* search, it is compared to a human summary using an evaluation metric such as ROUGE [29]. To learn the feature weights we used the training set (see Section VI-A) and generated for each aspect a list of candidate summaries (100 summaries were created for each topic). The summaries along with their feature values and ROUGE scores are input to a MERT (Minimum Error Rate Training) module to train the weights. MERT is a first order optimization method using Powell search to find the parameters which minimize the loss on the training data [30]. It is commonly used for training statistical machine translation systems.

V. Feature Extraction: Ratings Prediction Model

In order to apply the search techniques described above, it is necessary to define a set of features relevant to the summarization task that can be determined at each step of the search process.

A. Rating prediction model training

As previously mentioned, reviews refer to specific aspects of the product or service. For instance, reviews about restaurants will express opinions about the quality of the food, the courtesy of the wait personnel, or the ambience. These aspects are typically rated with a certain number of stars ranging from one (poor) to five (excellent). Based on [22], it is possible to create a rating prediction model that, for each aspect $a_i \in \mathcal{A}$, (e.g., $\mathcal{A} = \{food, service, ambience, value, overall\}$), estimates the ratings $r_i \in \mathcal{R}$ (e.g., $\mathcal{R} = \{1, \ldots, 5\}$) for any review document d_j in the considered document corpus $d_j \in \mathcal{D}$, as:

$$\hat{r_i} = \arg\max_{r \in \mathcal{R}} P(r_i | d_j) \quad (5)$$

$$= \arg\max_{r \in \mathcal{R}} P(r_i | s_{1,j}, s_{2,j}, \ldots, s_{n,j}) \quad (6)$$

where each document d_j is composed of n sentences or phrases $s_{1,j}, s_{2,j}, \ldots, s_{n,j}$. We used a maximum entropy (MaxEnt) [31] model to estimate the conditional probability of the ratings (Eq. 6) given the features extracted from the text documents.

Figure 2 shows the architecture of our rating prediction model system. In this configuration, each review document is associated with a set of predefined aspects that have been assessed by the reviewers with star-rating evaluations. During training, text features such as n-grams, parts of speech, shallow parsing chunks and others are used together with the reviewer-assigned ratings to create a discriminative

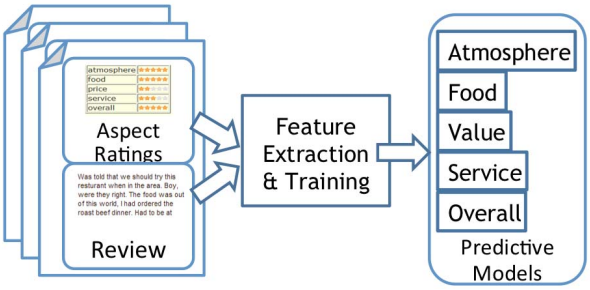

Figure 2. Rating predictive models

model classifier. There are five classifiers, one for each aspect.

For each document, the MaxEnt models will produce an estimate of the rating probability distributions $\hat{r}_i$, describing how ratings, for a specific aspect, are likely to be associated to the document text. For instance, a review like the following: *The service was flawless timely and non intrusive. Everything was great.* would have a rating distribution for the aspects *service* and *overall* skewed toward the high ratings (four or five) and almost uniform rating distribution for the remaining aspects.

The rating prediction models are trained from 6,823 restaurant reviews using the approach described in [22]. Performances measured in terms of rank loss average 0.63 across all the aspects.

B. Features extraction

The feature extraction process starts by calculating a set of scores for the initial candidate sentences. Although the rating prediction model was trained with labels associated to the whole review documents, we assume that model generalizes to the single sentences providing accurate probability distributions over the ratings. For each sentence and each aspect, it calculates the KL-divergence [32] between the predicted sentence rating distribution (as provided by MaxEnt) and a target rating distribution. The target rating distribution is simply calculated by numerically aggregating the ratings of the input reviews. The KL-divergence will quantify how far the current distribution is from the target. For each restaurant in the training set, we create a reference rating distribution by counting the rating contributions for each aspect in every review. The assumption is that the target text summary for a restaurant should reflect the same reference rating distribution calculated on the whole reviews set, but with many fewer words. In other words, the sentences should be selected based on their contribution toward the target rating distribution. The iterative process continues as described in Section IV till the MERT weights converge. Once the training ends, the optimal λ coefficients are used to in the A* search heuristic to traverse the graph and generate the optimal summary from the test set.

Table I
TEST DATA SET (20 RESTAURANTS) VALUES PER DOCUMENT SET

	Min	Max	Avg	Total
Reviews	6	10	7.55	151
Sentences	15	140	54.4	1,088
Words	206	2,042	809.85	16,197

Table II
TRAIN DATA SET (40 RESTAURANTS) VALUES PER DOCUMENT SET

	Min	Max	Avg	Total
Reviews	6	10	7.5	300
Sentences	15	108	51.95	2,078
Words	205	1,902	789.95	31,598

VI. EXPERIMENTAL SETUP

A. Data

The review documents used in our experiments were selected from a corpus of previously mined restaurant reviews from the we8there.com web site. In addition to the textual data, we8there.com provides numerical ratings for five predefined aspects: *atmosphere, food, value, service*, and *overall*. From the set of 3,866 available restaurants, we selected 131 with more than five reviews. Then, we manually searched for extra reviews on other web sites and selected 60 of the 131 restaurants that had reviews highly voted by web readers as useful. For each of the 60 restaurants, we selected the review with the highest number of "helpful votes" that was dated in the same time frame as the we8there.com reviews and use it as a reference summary. We randomly split the 60 restaurants into 40 for training and 20 for testing. Tables I and II gives info about the two data sets.

B. Entire Process

Our summarization process starts by calculating a set of features for every sentence in the training and testing set as described in Section V. In these experiments, the features we optimized with regard to the ROUGE score are the KL-divergence measure between the target aspect rating distribution and the predicted rating distributions calculated on the input sentences. The assumption is that the resulting summary will cover the target sentiments expressed in term of star-ratings by selecting the content that closer mimics the desired distribution and, at the same time, remains within the maximum summary size (100 words). In the next step, the features are weighted and their summation is used to score each sentence. The weights are trained using the training set and the algorithms described in Section IV. Finally, the trained weights are used in the scoring functions used by A* search to generate summaries for the test set.

VII. RESULTS

To evaluate our STARLET approach, we compared it with two summarizers: 1) a baseline summarization system that

randomly selects sentences with no repetition till it reaches the desired length of 100 words; 2) the open source MEAD system with the same output length. The resulting summaries were assessed using the automatic metric ROUGE and manual evaluation. Examples summaries are shown in Table III.

Random Summary

We ended up waiting 45 minutes for a table 15 minutes for a waitress and by that time they had sold out of fish fry s .
This would be at least 4 visits in the last three years and the last visit was in March 2004 .
During a recent business trip I ate at the Fireside Inn 3 times the food was so good I did n't care to try anyplace else .
I always enjoy meetiing friends here when I am in town .
The food especially pasta calabria is delicious .
I like eating at a resturant where I can not see the plate when my entry is served .

MEAD Summary

During a recent business trip I ate at the Fireside Inn 3 times the food was so good I did n't care to try anyplace else .
I have had the pleasure to visit the Fireside on every trip I make to the Buffalo area .
The Fireside not only has great food it is one of the most comfortable places we have seen in a long time The service was as good as the meal from the time we walked in to the time we left we could have not had a better experience We most certainly will be back many times .

STARLET Summary

Delicious .
Can't wait for my next trip to Buffalo .
GREAT WINGS .
I have reorranged business trips 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 service was not rushed and was very timely .
The food especially pasta calabria is delicious .
2 thumbs UP .
A great night for all .
the food is very good and well presented .
The price is more than competivite .
It took 30 minutes to get our orders .

Table III
EXAMPLE OF SUMMARIES EXTRACTED FROM A SET OF 10 RESTAURANT REVIEWS

A. ROUGE Evaluation

ROUGE is a well-known evaluation method for summarization, which is based on the common number of n-grams between a peer, and one or several model summaries. The metrics taken into consideration for this evaluation are ROUGE-1 (R-1), ROUGE-2 (R-2) and ROUGE-SU4 (R-SU4). R-1 and R-2 compute the number of unigrams and bigrams, respectively, that coincides in the automatic and model summaries. R-SU4 measures the overlap of skip-bigrams between them allowing a skip distance of 4.

Table IV
ROUGE SCORES OBTAINED FROM THE TEST SET

Metric	Random	MEAD	STARLET
R-1	0.2769	0.2603	0.2894
R-2	0.0329	0.0377	0.0454
R-SU4	0.0790	0.0727	0.0881

Table V
MANUAL EVALUATION FOR THE THREE SUMMARIZATION SYSTEMS

	Random	MEAD	*Starlet*
Grammatically	3.53	3.68	3.67
Redundancy	2.82	2.92	3.00
Clarity	2.78	2.97	3.05
Coverage	2.67	2.33	3.23
Coherence	2.05	2.57	2.62

From Table IV we can see that STARLET outperforms the other two systems in all ROUGE metrics. This means that, according to ROUGE, our summarizer generates reviews/summaries whose lexical content is closer to human ones and thus are more likely to capture the opinions about the restaurant than the other two systems.

B. Manual Evaluation

In the manual evaluation we asked three people (two of them native speakers of English) to evaluate the readability of the generated summaries according to the evaluation criteria described in [33]. Without showing the reference summary, we asked each participant to rate the following linguistic qualities with a rating scale ranging from a maximum of 5 (very good) to a minimum of 1 (very poor): **Grammaticality** - grammatically correct and without artifacts; **Redundancy** - absence of unnecessary repetitions; **Clarity** - easy to read; **Coherence** - well structured and organized. Since the **Focus** readability property listed in [33] applies mostly to the DUC summarization tasks, we replaced it with **Coverage** that indicated level of coverage for the aspects and the polarity expressed in the summary. In other words, this rating should be higher if most of the sentences are expressing opinions on the pre-defined restaurant aspects. The average scores for each criterion are shown in Table V.

From Table V we can see that the scores for *Grammaticality, Redundancy, Clarity* and *Coherence* are in all systems very close to each other. The only gap can be observed in the *Coverage* metric. This metric expresses how many opinions and aspects are actually covered in the review/summary. The

scores indicate that STARLET is able to generate summaries with a wider range of opinions than the other two systems.

C. Discussion

Reviews are typically written by consumers to convey their personal opinion of a product or a service. Compared to traditional automatic summarization tasks – where the documents are usually written by professionals, edited, and proofread – the English quality of reviews is usually poor, often ungrammatical, incoherent, and inconsistent. While in news-based documents, facts and events are the central topics expressed by the author, reviews are focused on attributes or features of a product or service and the reviewer's opinions about the qualities of such characteristics.

Looking at the manual evaluation from the judges, the 'grammatically' scores are consistent across the three methods and depend only on the quality of the source sentence. Poorly written sentences can be penalized by introducing new features during training that take into consideration the number of misspellings (for instance, in our data, the word *atmosphere* has been misspelled in 23 different ways), the number of words belonging to the English dictionary, and scores from a parser.

The 'redundancy' score is slightly better for STARLET, but in the current version there is not a mechanism to avoid similar sentences, although selecting sentences according to the rating distribution should help to reduce redundancy. Sentence similarity features can be added during training by using centroid-based clustering and demote similar sentences to these already included in the summary.

Also the 'clarity' and 'coherence' scores are better in our approach. Low scores are related to controversial reviews where the opinions are mixed and distributed across the ratings. In these cases, more investigation is necessary, perhaps ordering positive and negative sentences according to some rhetorical structure or learned-from-data language models.

Finally, the 'coverage' score for STARLET is decidedly better than for the other approaches, showing that STARLET correctly selects information relevant to the users.

VIII. Conclusions and Future Work

This paper addresses extractive summarization for reviews containing opinions on multiple aspects of the product or serivice being reviewed. We propose a method called STARLET. It uses aspects as features to score sentences in the input documents. The features are weighted linearly and summaries are generated using A* search. We trained the weights using MERT and use "best" reviews as gold standard summaries. We performed both automatic and manual evaluations in the restaurant reviews domain. In both evaluations the results show that STARLET summaries contain more review information than alternative baselines.

In future work we plan to study the integration of other aspects such as redundancy and coherence into the feature weight learning process. In addition to this we plan to investigate more appropriate evaluation metrics for review summaries and explore more features as described in the discussion section.

Acknowledgment

The authors would like to thank Amanda Stent, Alistair Conkie, Patrick Haffner, Narendra Gupta, and the anonymous reviewers who provided useful feedback and suggestions.

References

[1] W. Duan, B. Gu, and A. B. Whinston, "Do Online Reviews Matter? - An Empirical Investigation of Panel Data," *Journal Decision Support Systems*, vol. 45, no. 4, pp. 1007–1016, November 2008.

[2] D. Park, J. Lee, and I. Han, "The effect of on-line consumer reviews on consumer purchasing intention: The moderating role of involvement," *Int. J. Electron. Commerce*, vol. 11, pp. 125–148, July 2007.

[3] J. Feng, M. Johnston, and S. Bangalore, "Speech and multimodal interaction in mobile search," *Signal Processing Magazine, IEEE*, vol. 28, no. 4, pp. 40 –49, july 2011.

[4] A. Stent, I. Zeljković, D. Caseiro, and J. Wilpon, "Geocentric language models for local business voice search," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, ser. NAACL '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 389–396.

[5] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008.

[6] R. McDonald, K. Hannan, T. Neylon, M. Wells, and J. Reynar, "Structured models for fine-to-coarse sentiment analysis," in *Proceedings of the Association for Computational Linguistics (ACL)*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 432–439.

[7] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2004, pp. 168–177.

[8] J. M. Conroy, J. G. Stewart, and J. D. Schlesinger, "CLASSY query-based multi-document summarization," in *In Proceedings of the Document Understanding Conf. Wksp. 2005 (DUC 2005) at the Human Language Technology Conf./Conf. on Empirical Methods in Natural Language Processing (HLT/EMNLP*, 2005.

[9] K. R. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, J. L. Klavans, A. Nenkova, C. Sable, B. Schiffman, and S. Sigelman, "Tracking and summarizing news on a daily basis with columbia's newsblaster," in *Proceedings of the second international conference on Human Language Technology Research*, ser. HLT '02. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2002, pp. 280–285.

[10] N. Elhadad, M. Y. Kan, J. L. Klavans, and K. R. McKeown, "Customization in a unified framework for summarizing medical literature," *Artif. Intell. Med.*, vol. 33, pp. 179–198, February 2005.

[11] T. Copeck, N. Japkowicz, and S. Szpakowicz, "Text summarization as controlled search," in *Proceedings of the 15th Conference of the Canadian Society for Computational Studies of Intelligence on Advances in Artificial Intelligence*, ser. AI '02. London, UK, UK: Springer-Verlag, 2002, pp. 268–280.

[12] H. Saggion and L. Guy, "Generating indicative-informative summaries with sumum," *Comput. Linguist.*, vol. 28, pp. 497–526, December 2002.

[13] S. Mithun and L. Kosseim, "Summarizing blog entries versus news texts," in *Proceedings of the Workshop on Events in Emerging Text Types*, ser. eETTs '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 1–8.

[14] L. W. Ku, Y. T. Liang, and H. H. Chen, "Opinion extraction, summarization and tracking in news and blog corpora," in *Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs*, 2006.

[15] A. Aker, T. Cohn, and R. Gaizauskas, "Multi-document summarization using a* search and discriminative training," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2010, pp. 482–491.

[16] I. Titov and R. McDonald, "A joint model of text and aspect ratings for sentiment summarization," in *Proceedings of ACL-08: HLT*, 2008, pp. 308–316.

[17] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai, "Topic sentiment mixture: modeling facets and opinions in weblogs," in *WWW '07: Proceedings of the 16th international conference on World Wide Web*. ACM, 2007, pp. 171–180.

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

[19] A. Popescu, B. Nguyen, and O. Etzioni, "OPINE: Extracting product features and opinions from reviews," in *Proceedings of HLT/EMNLP 2005 Interactive Demonstrations*. Association for Computational Linguistics, 2005, pp. 32–33.

[20] G. Qiu, B. Liu, J. Bu, and C. Chen, "Opinion word expansion and target extraction through double propagation," *Comput. Linguist.*, vol. 37, pp. 9–27, 2011.

[21] A. Esuli, "Automatic generation of lexical resources for opinion mining: models, algorithms and applications," *SIGIR Forum*, vol. 42, pp. 105–106, November 2008.

[22] N. Gupta, G. Di Fabbrizio, and P. Haffner, "Capturing the stars: predicting ratings for service and product reviews," in *Proceedings of the NAACL HLT 2010 Workshop on Semantic Search*, ser. SS '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 36–43.

[23] B. Snyder and R. Barzilay, "Multiple aspect ranking using the good grief algorithm," in *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*. Association for Computational Linguistics, 2007, pp. 300–307.

[24] S. Baccianella, A. Esuli, and F. Sebastiani, "Multi-facet rating of product reviews," in *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, ser. ECIR '09. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 461–472.

[25] S. Blair-Goldensohn, K. Hannan, R. McDonald, T. Neylon, G. Reis, and J. Reynar, "Building a Sentiment Summarizer for Local Service Reviews," in *NLPIX*, 2008.

[26] G. Carenini and L. Rizoli, "A multimedia interface for facilitating comparisons of opinions," in *IUI '09: Proceedings of the 13th international conference on Intelligent user interfaces*. ACM, 2009, pp. 325–334.

[27] G. Carenini, R. Ng, and A. Pauls, "Multi-document summarization of evaluative text," in *11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, 2006.

[28] D. Radev, T. Allison, S. Blair-Goldensohn, J. Blitzer, A. Çelebi, S. Dimitrov, E. Drabek, A. Hakim, W. Lam, D. Liu, J. Otterbacher, H. Qi, H. Saggion, S. Teufel, M. Topper, A. Winkel, and Z. Zhang, "MEAD — A platform for multidocument multilingual text summarization," in *Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal, May 2004.

[29] C. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proceedings of the workshop on text summarization branches out (WAS 2004)*, 2004, pp. 25–26.

[30] F. Och, "Minimum error rate training in statistical machine translation," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, 2003, pp. 160–167.

[31] A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra, "A maximum entropy approach to natural language processing," *Comput. Linguist.*, vol. 22, pp. 39–71, March 1996.

[32] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.

[33] H. Dang, "Overview of DUC 2005," *DUC 05 Workshop at HLT/EMNLP*, 2005.