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

Giuseppe “Pino” Di Fabbrizio*, Ahmet Aker ^, and Rob Gaizauskas ^

ICDM 2011 – SENTIRE Workshop – Vancouver, CA
December 11, 2011
Outline

• Introduction
• Summarization as search problem
  ▪ A* search
  ▪ Feature extraction
  ▪ Star rating prediction model
  ▪ Training
• Experiments
• Results and discussion
Questions

• Summarization - What does it mean to summarize reviews?
• Star ratings – Does the number of star provide enough information?
• Selection process – What is important to preserve?
• Learning from data – Can we learn what is relevant from data?
• Controversiality – What do we do about contradictory information?
A reasonable goal

- Given a set of reviews evaluating a specific entity (restaurant, hotel, digital camera, etc.) and related aspects describing the entity (food, service, atmosphere, etc.)

   ✷ Extract the sentences with relevant information about the evaluated aspects preserving the average opinions distributions

<table>
<thead>
<tr>
<th>N</th>
<th>Review 931–5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Rude employees</td>
</tr>
<tr>
<td>2</td>
<td>Bartenders are the worst</td>
</tr>
<tr>
<td>3</td>
<td>An extremely local hang out</td>
</tr>
<tr>
<td>4</td>
<td>If not a friend of the crew be prepared to wait and no friendly attitudes</td>
</tr>
<tr>
<td>5</td>
<td>Bar top a mess and always wet</td>
</tr>
<tr>
<td>6</td>
<td>Best thing is the T.V.'s showing sports</td>
</tr>
<tr>
<td>7</td>
<td>Live music there is ok not great</td>
</tr>
<tr>
<td>8</td>
<td>Some nice decor and there are pool tables with room to play</td>
</tr>
<tr>
<td>9</td>
<td>More for the day crowd</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Aspects</th>
<th>Ratings</th>
<th>Stars</th>
</tr>
</thead>
<tbody>
<tr>
<td>atmosphere</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>food</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>overall</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>price</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>service</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>N</th>
<th>Review 931–4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Not a place to go for dinner</td>
</tr>
<tr>
<td>2</td>
<td>This is the type of place you go for live music reggae punk ska live sound system</td>
</tr>
</tbody>
</table>

Dec 11, 2011
Automatic summarization

The process of distilling the most important information from a **text** to produce an **abridged** version for a particular task and users.

[Mani and MayBury, 1999]

- **Methods**
  - Extractive – text units (phrase / sentence) selection
  - Compression – text simplification
  - Abstractive – natural language generation

- **Evaluation metrics**
  - Intrinsic – human generated (gold) reference
  - Extrinsic – evaluated according some utility function (i.e., document snippet accuracy in web search)

- **Input / Output**
  - Text, speech, graphics (any combination)
Multi-document summarization

- Traditional multi-document summarization (DUC, TAC)
  - Focuses on facts, usually coherent and non contradictory
  - Edited, high quality written text
  - Limited number of documents (<<100)
  - Typical approach
    - Sentences clustering, selection, and ordering in a domain-independent way
Typical summarization tasks

- News articles
  - [McKeown et al., 2002]
- Medical literature
  - [Elhadad et al., 2005]
- Biographies
  - [Copeck et al., 2002]
- Technical articles
  - [Saggion and Guy, 2001]
- Blogs
  - [Mithhun and Kosseim, 2009]
Multi-document summarization (opinion)

- Multi-document summarization for **evaluative text**
  - Contradictory opinions
  - Poorly written (typos, misspellings, ungrammatical, jargon)
    - 20 different ways to misspell *atmosphere*:
      atmosphere, atmopshere, atmoshere, atmosphere, atmosphere, atmosphere, atmospphere, atmospshere, atmospere, atmosphere, atmosphere, atmoslhere, atmospheric, atmosphere, atmosphere, atmosphere, atmosphere, atmosphere, atmosphere, atomosphere, atomosphere, atomosphere, atomosphere, atomosphere, atomosphere, atomosphere, atomosphere, atomosphere, atomosphere, atomosphere, atomosphere, atomosphere, atomosphere
  - Vast range of domains (restaurants, hotels, cars, books, toasters, etc.)
  - Number of documents could be large for popular products (>200)
- Typical approach
  - Sentence selection on sentiment-laden sentences
  - Template-based natural language generation
MEAD*
[Carenini et al., 2006, Carenini et al. 2011]

• Based on MEAD [Radev et al., 2003], an open source, PERL-based extractive summarizer

• Three steps process
  ▪ Feature calculation – evaluate how informative is the sentence. Use centroids and evaluative features
  ▪ Classification – combine features in one score
  ▪ Reranking – sentence scores adjustments based on the number of opinions present in a sentence (regardless of the polarity)

• Drawbacks
  ▪ Sentence selection based on most frequently discussed aspects
  ▪ Polarity of sentences is ignored (positive and negative sentences have the same contribution)
  ▪ Summarization features based on expert knowledge
Summarization as search problem

- Scoring function as linear combination of summarization features

\[ s(y|x) = \Phi(y|x; \lambda) \]

where

- \( x \) is a vector of indexes representing the \( N \) sentences in the document set to summarize

- \( y \subseteq \{1, \ldots, N\} \) is the set of indexes selected for the summary of length \( |y| = M \)

- \( \lambda = \{\lambda_1, \ldots, \lambda_F\} \) is the weight vector of parameters for the \( F \) features that optimizes the summary evaluation metrics

- \( \Phi(\cdot|\cdot) \) is a function that returns a set of features for each candidate summary
Summarization model

- Assuming that the features are independent

\[ s(y|x) = \sum_{i \in y} \phi(x_i) \lambda_i \]

- Find the parameters \( \lambda_i \) such that \( \hat{y} \) score is similar to the score from a gold standard summary

\[ \hat{y} = \arg \max_y s(y|x) \]

- Exponentially large search space

\[ O(S^{L(W)}) \]

- where \( S \) is the total number of sentences and \( L(W) \) is the number of sentences that best matches the required summary word length \( W \)
Goal

*Find the best scoring path from $S$ to $E*$

Reviews

Sentences

Summaries

[S] Aker et al., 2010

Dec 11, 2011
A* search

- Sooo many stars ...
- Informed search algorithm
- Best-first strategy

Guarantee to find optimal solution if heuristic function is **monotonic** or follows the **admissible heuristic** requirement:

- Estimated cost from the current node to the goal node never overestimates the actual cost
- For the node $n$: $f(n) = s(n) + h(n)$
- Where
  - $s(n)$ - sum of the current scores based on the summary so far
  - $h(n)$ - heuristic function to estimate how far from the final summary length [Aker et al., 2010]

- Heuristic keeps in consideration global constraints such as ‘summary length’
Model parameter optimization

- Find the parameters $\lambda_i$ such that $\hat{y}$ score is similar to the score from a gold standard summary

$$\hat{y} = \arg \max_y s(y|x)$$

- ROUGE metric to measure accuracy of the current summary $\hat{y}$ with a gold reference summary $r$

- Minimize the loss function

$$\hat{\lambda} = \arg \min_\lambda \Delta(\hat{y}|r)$$

- Minimum error rate training (MERT) [Och, 2003]
- First order approximation method using Powell search (not convex)
- Iterative method, uses n-best candidates in A* search to find parameters
Feature extraction

- Rating prediction model

For each aspect \( a_i \in \{ \text{food, service, ambience, value, overall} \} \) estimate the ratings \( r_i \in \{1, \ldots, 5\} \) for any document \( d_j \in \mathcal{D} \)

\[
\hat{r}_i = \arg \max_{r \in \mathcal{R}} P(r_i|d_j) \quad (1)
\]

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

- MaxEnt classification algorithm trained on 6,823 restaurant reviews with an average rank loss of 0.63

- Predicts rating distributions (after proper confidence score normalization)
Predicted and target ratings

Dec 11, 2011

pino

Predicted food ratings

Average food ratings
Review ratings as summarization features

- For each review document set
  - For each aspect $i$, average the ratings by aspect to create target reference distribution $\bar{r}_i$
  - For each sentence $j$, calculate aspect rating predictions $\hat{r}_{i,j}$
  - For each sentence, calculate Kullback–Leibler divergence with the reference summary
    \[ D_{KL}^{i,j}(\hat{r}_{i,j} \mid \bar{r}_i) \]
- KL-divergence is used then used during training to find optimal parameters
Data

- From 3,866 available restaurants (we8there.com), selected 131 with more than five reviews
- Selected 60 over 131 restaurants that had reviews on tripadvisor.com highly voted by by readers as useful
- Created the GOLD reference by selecting the 20 reviews from tripadvisor.com with the highest number of “helpful votes” (same time frame as the we8there.com reviews)
- Remaining 40 restaurants used as training set

<table>
<thead>
<tr>
<th>Table I</th>
<th>Test data set (20 restaurants) values per document set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Min</td>
</tr>
<tr>
<td>Reviews</td>
<td>6</td>
</tr>
<tr>
<td>Sentences</td>
<td>15</td>
</tr>
<tr>
<td>Words</td>
<td>206</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table II</th>
<th>Train data set (40 restaurants) values per document set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Min</td>
</tr>
<tr>
<td>Reviews</td>
<td>6</td>
</tr>
<tr>
<td>Sentences</td>
<td>15</td>
</tr>
<tr>
<td>Words</td>
<td>205</td>
</tr>
</tbody>
</table>
Experimental setup

- Target length: 100 words
- Baseline
  - Randomly selected sentences with no repetition till it reaches the target length
- MEAD
  - Traditional multi-document summarization
- Starlet
  - Using only rating distributions as feature and web-based GOLD reference
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 meeting 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 reorarranged 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.

2 thumbs UP.
A great night for all.
The food is very good and well presented.
The price is more than competitve.

It took 30 minutes to get our orders.
ROUGE evaluation

<table>
<thead>
<tr>
<th>Metric</th>
<th>Random</th>
<th>MEAD</th>
<th>STARLET</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-1</td>
<td>0.2769</td>
<td>0.2603</td>
<td>0.2894</td>
</tr>
<tr>
<td>R-2</td>
<td>0.0329</td>
<td>0.0377</td>
<td>0.0454</td>
</tr>
<tr>
<td>R-SU4</td>
<td>0.0790</td>
<td>0.0727</td>
<td>0.0881</td>
</tr>
</tbody>
</table>
Manual evaluation

- Three judges (two native speakers)
- Rating scale: 5 (very good) to 1 (very poor)
- Evaluations
  - Grammaticality - grammatically correct and without artifacts
  - Redundancy - absence of unnecessary repetitions;
  - Clarity - easy to read
  - Coverage - level of coverage for the aspects and the polarity expressed in the summary
  - Coherence - well structured and organized

<table>
<thead>
<tr>
<th></th>
<th>Random</th>
<th>MEAD</th>
<th>Starlet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammatically</td>
<td>3.53</td>
<td>3.68</td>
<td>3.67</td>
</tr>
<tr>
<td>Redundancy</td>
<td>2.82</td>
<td>2.92</td>
<td>3.00</td>
</tr>
<tr>
<td>Clarity</td>
<td>2.78</td>
<td>2.97</td>
<td>3.05</td>
</tr>
<tr>
<td>Coverage</td>
<td>2.67</td>
<td>2.33</td>
<td>3.23</td>
</tr>
<tr>
<td>Coherence</td>
<td>2.05</td>
<td>2.57</td>
<td>2.62</td>
</tr>
</tbody>
</table>
Discussion

- **Grammatically** - 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
- **Redundancy** - slightly better for Starlet. Sentence similarity features can be added during training by using centroid-based clustering and demote similar sentences to these already included in the summary.
- **Clarity** and **coherence** - slightly better in Starlet, but more investigation is necessary
- **Coverage** - decidedly better than for the other approaches, showing that Starlet correctly selects information relevant to the users
Conclusions

- Summarization - What does it mean to summarize reviews?
- Star ratings – Does the number of star provide enough information?
- Selection process – What is important to preserve?
- Learning from data – Can we learn what is relevant from data?
- Controversiality – What do we do about contradictory information?