* at&t Labs - Research

^ The University Of Sheffield.

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

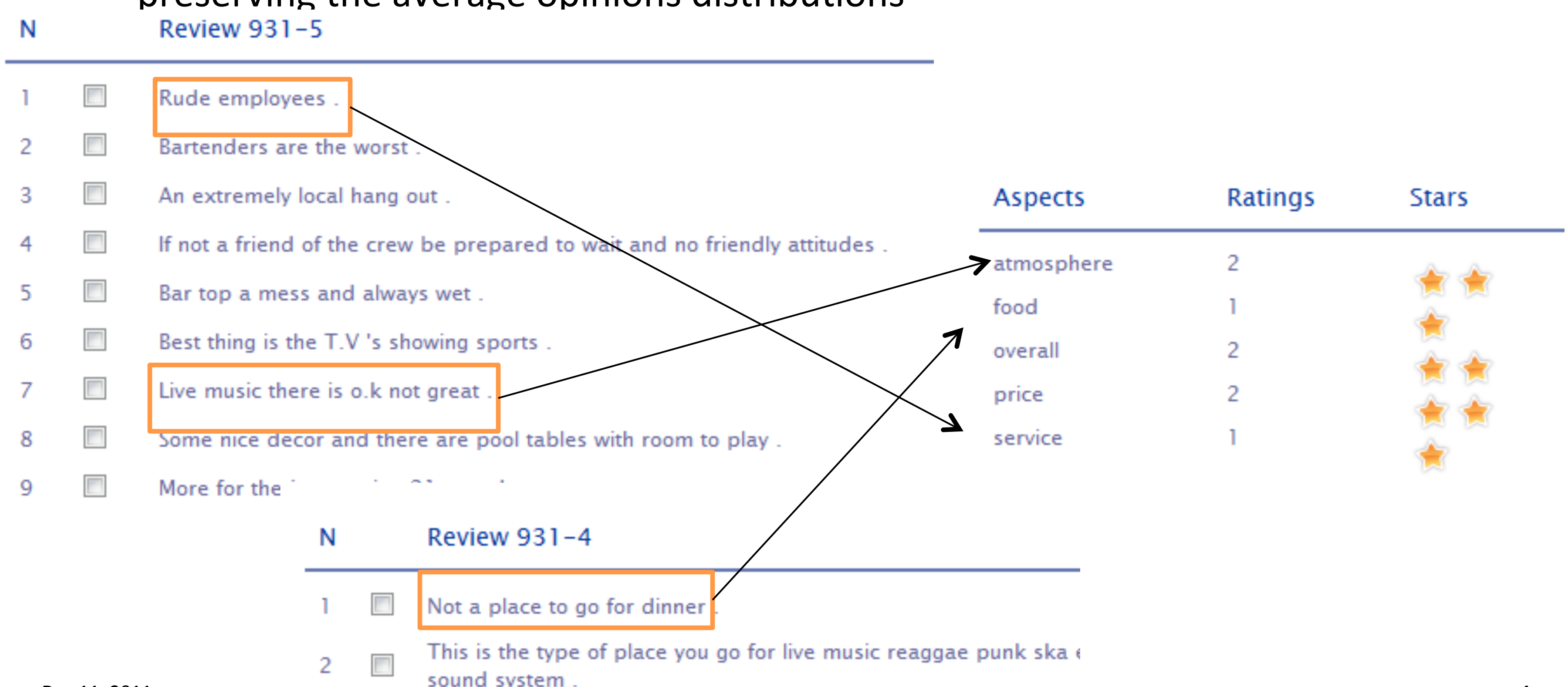
- Introduction
- Summarization as search problem
  - A* search
  - Feature extraction
  - Star rating prediction model
  - Training
- Experiments
- Results and discussion

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



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**:
      atmophere, atmopshere, atmoshere, atmoshpere, atmoshphere, atmosophere, atmospehere, atmospere, atmosphare, atmoslhere, atmospheric, atmosphire, atmosphre, atmostphere, atmousphere, atmsphere, atomosphere, atompospere, atomsphere, atsmosphere
  - 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

$$s(\mathbf{y}|\mathbf{x}) = \Phi(\mathbf{y}|\mathbf{x};\lambda)$$

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 $|\mathbf{y}| = M$
- $\lambda = \{\lambda_1,\ldots,\lambda_F\}$ 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

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

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

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

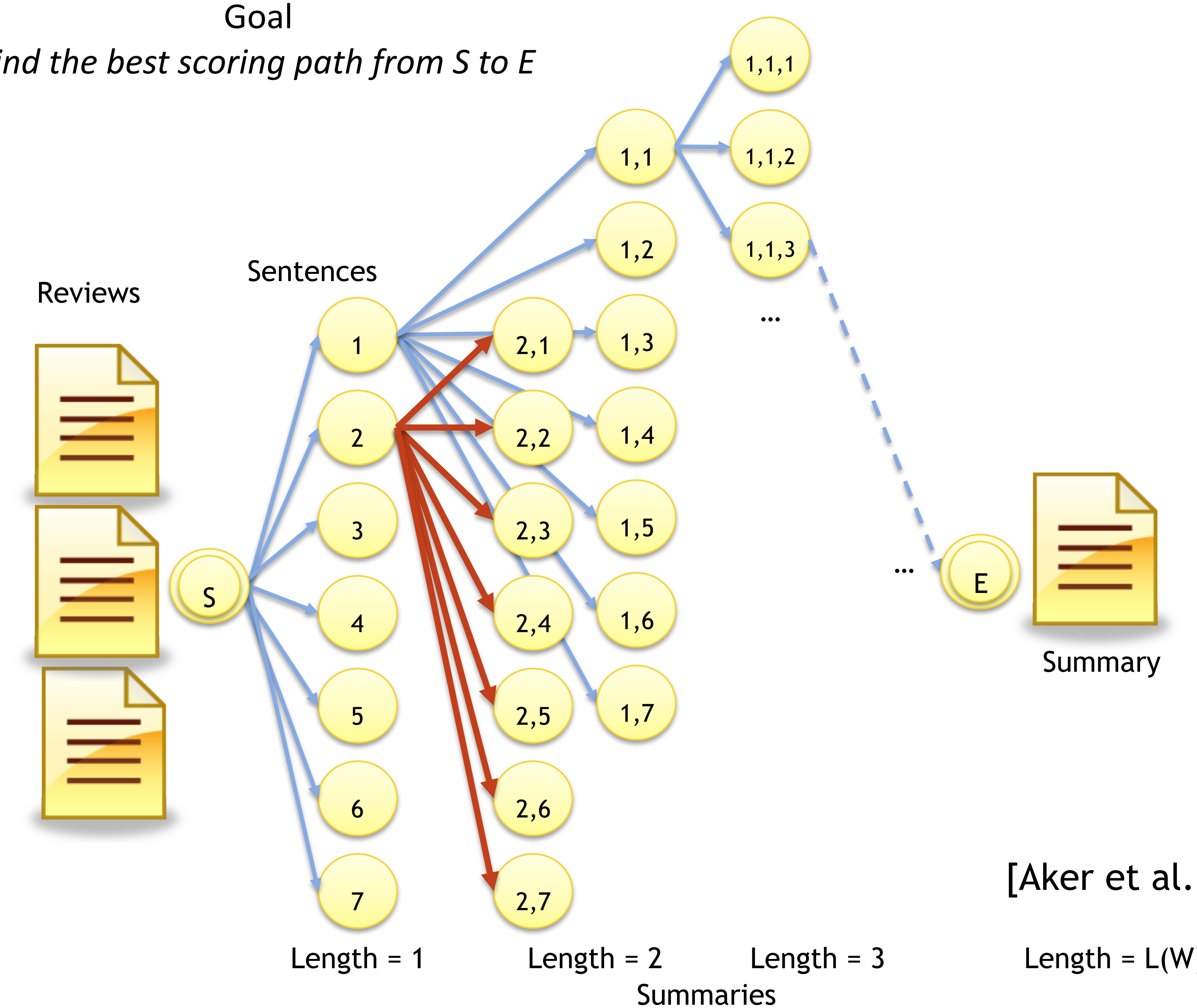
- Exponentially large search space

$$\mathcal{O}(S^{L(W)})$$

- 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



[Aker et al., 2010]

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 λ_i such that $\hat{\mathbf{y}}$ score is similar to the score from a gold standard summary

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

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

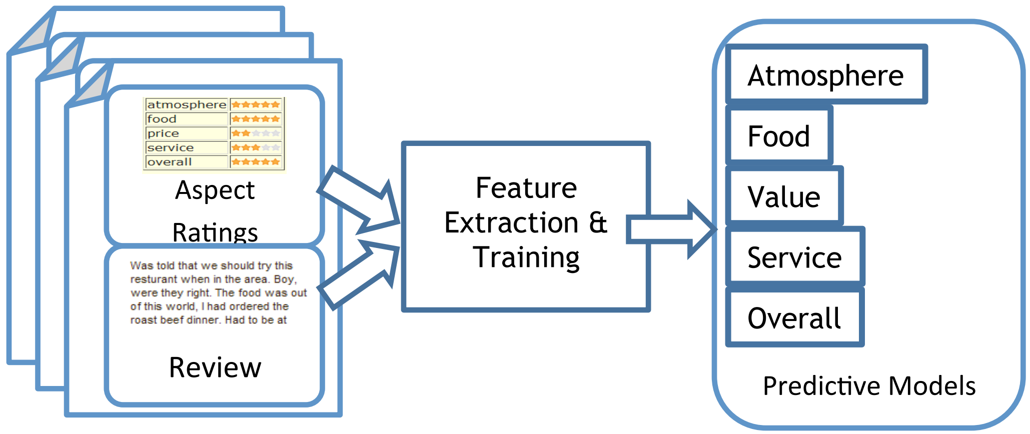
$$\hat{\lambda} = \arg\min_{\lambda} \Delta(\hat{\mathbf{y}}|\mathbf{r})$$

- 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

[Gupta, Di Fabbrizio, Haffner, 2010]

- Rating prediction model



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

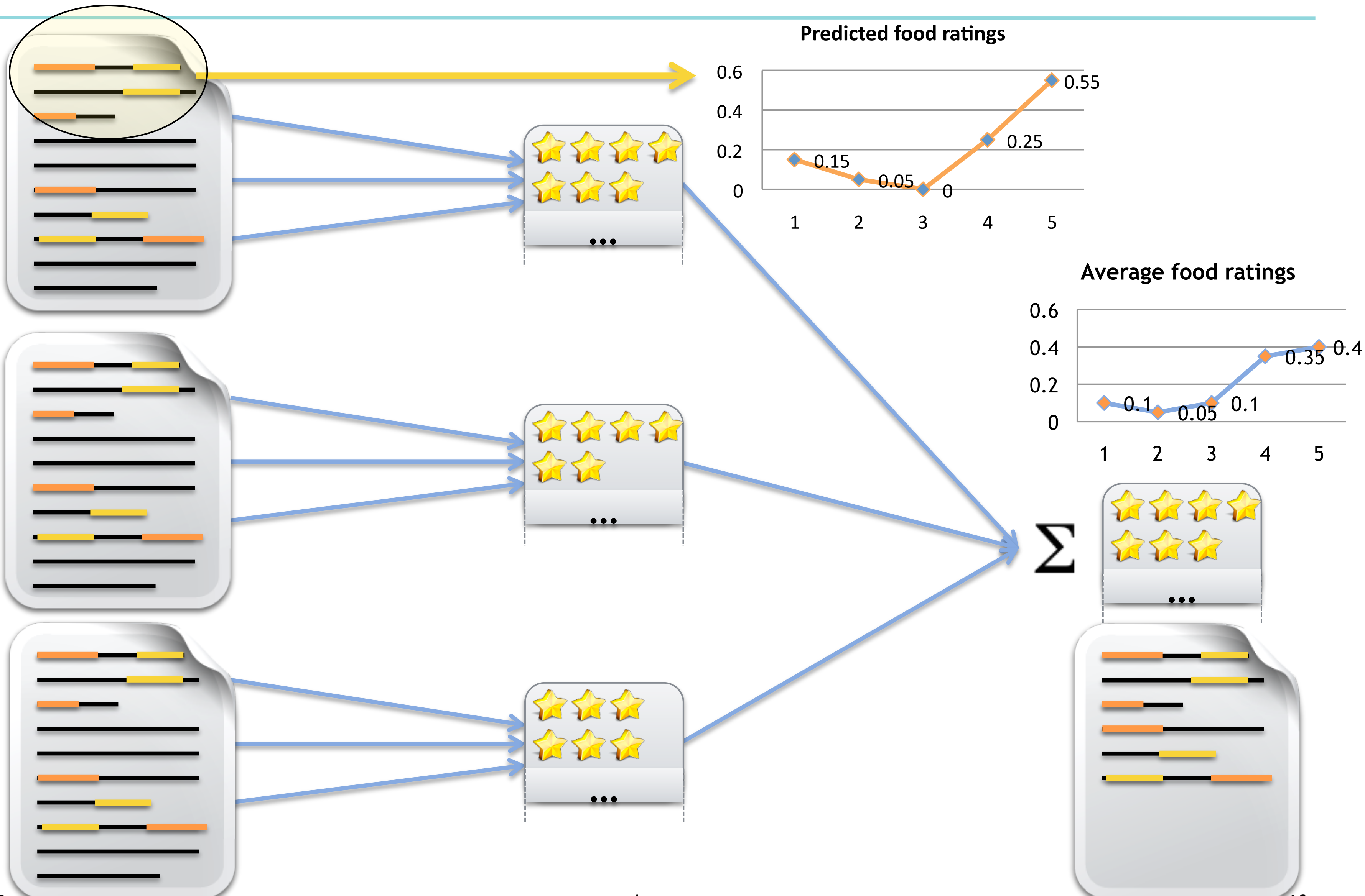
$$\hat{r_i} = \arg\max_{r \in \mathcal{R}} P(r_i | d_j) \quad (1)$$

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

- 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

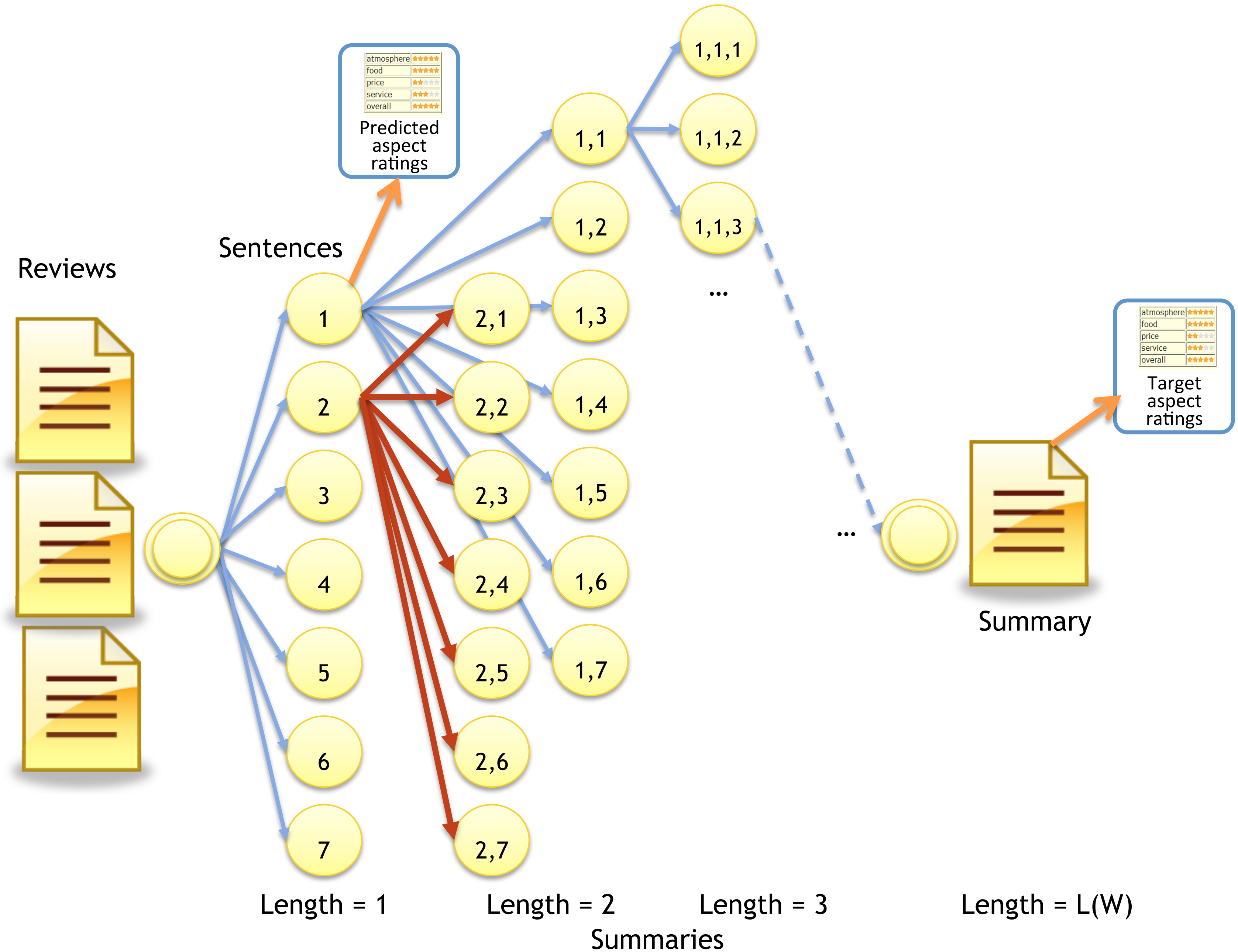


Review ratings as summarization features

- 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

$$D_{KL}^{i,j}(\hat{r}_{i,j}||\bar{r}_i)$$

- KL-divergence is used then used during training to find optimal parameters


Reviews
Sentences
atmosphere
food
price
service
overall
Predicted aspect ratings
1
2
3
4
5
6
7
1,1
1,2
1,3
1,4
1,5
1,6
1,7
2,1
2,2
2,3
2,4
2,5
2,6
2,7
1,1,1
1,1,2
1,1,3
...
...
atmosphere
food
price
service
overall
Target aspect ratings
Summary
Length = 1
Length = 2
Length = 3
Length = L(W)
Summaries

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 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

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

Output example

Random Summary

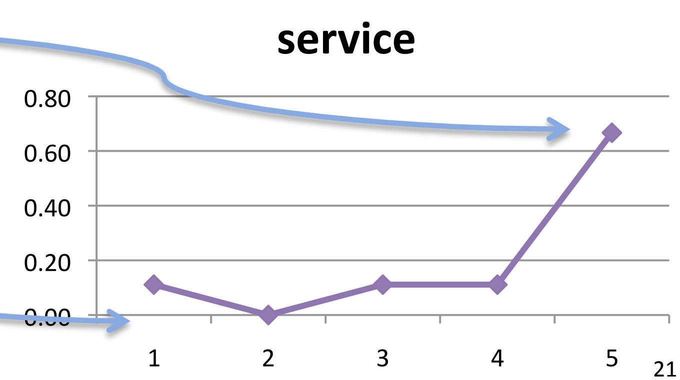
We ended up waiting 45 minutes for a table 15 minutes for a waitress and by that time they had sold out of fish fry s .
This would be at least 4 visits in the last three years and the last visit was in March 2004 .
During a recent business trip I ate at the Fireside Inn 3 times the food was so good I 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 competitive .
<u>It took 30 minutes to get our orders .</u>



ROUGE evaluation

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

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 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

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?