Summarization of Multiple User Reviews in the Restaurant Domain

Patrick Nguyen, Milind Mahajan and Geoffrey Zweig
{panguyen,milindm,gzweig}@microsoft.com

September 2007

Technical Report
MSR-TR-2007-126

Microsoft Research
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052
http://www.research.microsoft.com

1 Introduction

In recent years, web-based review systems have provided a valuable service to consumers by allowing them to share their assessments of goods and services, thus enabling more informed decision making. The use of these systems, however, requires access to a web browser – typically at home or at the office – and this restricts the usefulness of the systems to planned sorties. While mobile phones also provide web access, their small screens, flanked with unwieldy input modalities, severely limit their convenience. Thus, today's diners have no viable source of information to draw on while they are out and away from their desktop computers.

We have developed a voice rating telephone system for restaurants which fills this "information gap." By calling the system, prospective diners may inquire about potential meal venues. The system engages in a dialog with the user, locates the appropriate entry in its database, and plays the summary. Using voice interaction is natural and shortens the time to enter the name of the restaurant. We also offer options other than restaurants, such as products and services, but these are beyond the scope this paper.

For either voice output or rendering on a small screen, limitations in space make it necessary to compress the available information into a brief and compact answer. There may be potentially tens or hundreds of reviews for a particular establishment. The challenge which this paper aims to address is what would be fitting to tell the user. The techniques described herein are applicable to both screen and voice output, but we will assume the latter, as it is the current state of our prototype.

2 Overview

There are several possible ways of condensing review content. In this paper, we have explored only one of many. We think that our algorithm yields a satisfactory outcome, although we would not presume to suggest that other alternatives are invalid. We will first sketch the general goal with some potential desired characteristics. Then, according to those characteristics, we introduce the idea of extracting snippets from the database. Then, we continue to describe how the system works algorithmically.

2.1 Objectives

We set ourselves the goal of improving the voice rating system through the addition of specific differentiating information for restaurants. Ultimately, pieces of information may serve multiple purposes. Users consume information in different ways, and therefore there is no single tangible, measurable metric which encompasses all usage scenarios.

The first purported usage for review comments is to provide information about a restaurant. This is our primary goal. Ideally, one would render the most important facts or impressions about a restaurant. These, put together, would be the primary basic inputs which the user might use as a basis for his/her opinion about a restaurant, and ultimately, whether or not to visit the establishment. Therefore, it is a device for enriching the terse numeric rating.

A secondary goal is specific to automated computerized systems. There is an abundance of restaurant from which to choose, and the decision process is a human and somewhat emotional process. Review web sites address that issue by exposing users to other people's prose, so that an emotional connection with the community, and a sense of trust may be developed. That phenomenon is also a catalyst for contributing reviews. Presenting human reviews via our computerized system enhances the sense of confidence and credibility by exposing material which was underlies the overall rating.

The third kind of added value comes from keeping the user entertained. Adding randomness to the system, much the same way fortune cookies arouse interest, is at times beneficial to the system. Each restaurant review is distinguished by a specific if amusing signature – provided to us by the colorfulness of the reviewers.

2.2 Approach

Two main approaches prevail in summarization. In the first approach, one learns from and understands reviews for a given restaurant. Facts about a restaurant are gathered into an ontology system. They may be employed for summarization, and they also offer the flexibility of ready extension to other tasks (e.g. faceted search, or restaurant comparison). A summary would be *generated* from stored data. For instance, for restaurants, the ontology might cover food quality, atmosphere, and service: the generated review would cover each. In that case, we would be faced with the difficulty of extracting reliable facts, the complexity of the ontology, compounded with the challenge of incomplete data and other constraints such as legal requirements (involuntary libel). Therefore, this approach is hard to apply in our case. By contrast, in the second approach, one applies data driven approaches to *extract* a few exemplary snippets from a collection of reviews, e.g. [1]. The approach is easily portable to other tasks, but limited in scope, and does not usually result into more insight in the problem. Our approach is a hybrid of both approaches, e.g. [2, 3]. Snippet importance ranking is done with data-driven algorithms, but snippets are inscribed in an ontological category system for later extraction. The hope is that we may gather interesting insight which would later enable new services.

Thus, we set ourselves to the task of extracting relevant snippets from a review. In addition to finding such snippets, however, there is a secondary task: from the collection of candidate snippets to present from reviews, during a synthesis process, we then have to select and arrange those which will present a generally representative view of the restaurant. In the best case, we would do so in a way which conceals, or is robust to errors in the intermediate underlying process. Our approach decomposes the problem into two problems: a) finding snippets which are useful to render, and b) of all potentially interesting snippets to show to a user, given the limited time, determine which subset would, as a whole, give a good representative description of the restaurant.

There is no easy way of measuring the performance of the system beyond a mean opinion score (MOS) study. We will present intermediate metrics to measure the effectiveness of each subtask whenever possible.

2.3 System overview

The system overview is shown in Figure 1. As a preliminary step, we build a classifier which predicts, from each review, a numerical rating associated to the review. Then, each review is segmented into smaller units, called *snippets*. Ideally, these snippets would form self-contained, well-formed sentences; our implementation uses punctuation mark delimited sections instead. Each snippet is then assigned a bipolar relevance score, to mark it as characteristically disparaging (bad) or praising (good). Snippets are categorized into predefined types (such as food, or service). To produce the final review, we tally the most relevant snippets, both good and bad, in each category.

2.4 Database

We downloaded a database from a review site on the Internet. It consists of food establishments in urban areas in the United States of America. We crawled each of the 1600 listed urban centers for up to 2000 entries per urban center.

Summary statistics of the crawl are shown on Table 1. We left out 10k reviews for development, and 10k reviews for evaluation. To our knowledge, this is one of the largest scale databases used for review classification and summarization.

Examples of reviews are shown on Figure 2. Our qualitative analysis shows that the reviews tend to be self-contained and compact. Sometimes, they can be hard to follow as in the second case, but generally, they tend to be more of the well-formed type seen in the third case. Sentences tend to be connected, but still meaningful when read in isolation.

Number of reviewed restaurants	208,468
Number of reviews	326,146
Number of words	24,122,368
Avg number of words per review	73
Number of cities	13,915
Highly recommended (5)	57.22%
Recommended (4)	18.18%
Neutral (3)	8.92%
Below Average (2)	6.28%
Not Recommended (1)	9.40%

Table 1: General statistics about the web crawl.

3 Judging relevance

In summarizing reviews, we take the stance that one must first locate areas of high informational content. For this, we would like to extract the most extremely opinionated sentences. To do this, we look at the absolute magnitude of the contribution of each snippet to the overall rating of the review.

3.1 Log-linear models

Conditional log-linear models, also sometimes called *conditional maximum entropy* models are popular for natural language applications because of their simplicity and effectiveness. We use log-linear models for overall score regression, e.g. [4], and later for categorization.

3.1.1 Model

We would like to classify a review text r into several classes $\{c\}$. Choices for class assignment could range from 5 classes (Table 1) to two classes. Using finer classes allows for more features, and also provisions for the possibility that the "neutral" class would be distinctively different in itself, rather than a mid-point between good and bad. In other words, having five classes does not impose strict monotonicity constraints among the rating classes. On the other hand, this could lead to data sparsity and model estimation errors. Thus, there is a trade-off between using 5 classes and 2 classes. In preliminary experiments, however, the binary classifier (good vs bad) performed better according to the minimum squared error (MSE) criterion computed in a comparable fashion. We used a binary classifier in the remainder of this paper. For each class, we have a real-valued feature extraction function, $\mathbf{f}(r, c)$, of fixed dimension F. A log linear model with parameters λ defines a conditional probability of a class c given the review text r as:

$$p(c|r) = \frac{\exp\left[\lambda \cdot \mathbf{f}(r, c)\right]}{z(r; \lambda)}, \tag{1}$$

where $z(r; \lambda) = \sum_{c'} \exp[\lambda \cdot \mathbf{f}(r, c')]$.

3.1.2 Features

For simplicity, and to gain insight into the problem, we restricted ourselves to unigram and bigram features. A list of most frequent words was drawn from the training set. There were 40k words occurring more than 4 times, and 313k bigrams occurring more than 4 times. If a word w was seen $N(w, r)$ times in the review, we set its feature function to be:

$$f_w(r, c) = \log[1 + N(w, r)], \qquad \forall c. \tag{2}$$

Bigram features are treated similarly. In addition, a constant feature was added to account for class prior explicitly.

This bag of words approach might not be appropriate for some reviews. Although reviews tend to rate all aspects of a restaurants with approximately the same score, it is possible to have multiple conflicting comments within a review regarding the same restaurant. This phenomenon may occur at different levels of granularity. For instance, see Figure 3. Clearly, the feature extraction is limited and could benefit from segmentation and categorization of units. In practice, due to the concomitant effects of abundance of data, simple text structure, and uniformity within a review, we observed satisfactory results with our simple approach.

3.1.3 Objective and evaluation

We trained the log-linear model using the maximum *a posteriori* criterion over the training set, using Generalized Iterative Scaling (GIS) [5, 6]. We used a zero-mean normal prior with a global variance parameter tuned over the development set. The observations of 5 review classes in the training data were converted into binary class observations by using soft linear weights. For instance, for a review rated 3 in a 5 class rating, the binary class “good” was considered as observed with 0.75 soft weight and the other binary class “bad” was considered as observed with a soft weight of 0.25.

We evaluated the classifier with respect to its prediction performance for overall review ratings. Even though prediction of the overall review ratings is not our end-goal, it is an intermediate evaluation which allows us to judge the efficacy of the features and the models which are subsequently used for snippet selection. If we consider this as a regression task, a natural performance measure is the minimum mean squared error (MMSE) obtained by assigning numeric $1-5$ values to the 5 classes. We compare the MMSE of our classifier on the evaluation data with the prior distribution which assigns a constant rating value to each review without looking at the review text. The prior MMSE rating was 4.1, very close to a "Recommended" rating. The root mean squared (RMS) error for the prior was 1.30, and 0.67 for our classifier. If we remove the neutral ratings, we can evaluate the classification error rate for positive vs negative judgment: the error rate was 3.22%, much lower than other tasks such as movie reviews [4, 7].

Another commonly used evaluation measure for classifiers is the precision-recall trade-off. For example, in a task where we would like to retrieve the most interesting whole reviews, we would want high precision for the good and bad ratings. The ROC curve for precision-recall trade-off is shown on Figure 4.

3.2 Snippet extraction

Reviews are about 4 to 7 sentences in length. They need to be broken up into constituent snippets. Ideally, each snippet would comprise a single self-contained statement about the restaurant. Sentence breaking of largely ungrammatical text severely corrupted with misspellings is a hard problem. For simplicity and following a cursory analysis, we found that segmenting at all punctuation marks, including comma, yielded usable units. Keeping long units preserves context, but generally yields longer unwieldy snippets, and may also include more than one logical statement (e.g. food is good, but not the service). Having short units is more suited to aphoristic summarization. Striking the balance between those two extremes would be the subject of further research. We decide to err on over-segmentation, and hope that units without meaning will be pruned out in later stages. An example snippet decomposition of a review is given in Figure 5.

3.3 Snippet selection

In answer to the question of what would make a good snippet to present to users, we adopt the view that snippets which bear the most extreme views about a restaurant should be rendered. Furthermore, we assume that snippets whose words most contribute to the conditional log likelihood function are deemed important. This is distinct from selecting pieces of evidence most reflective of the review rating:

for an average restaurant, we do not select the more inane comments, but rather always try to pick the most contrasting views regardless of the score. The most prominent trigger features are shown in Figure 6. We apply our model, trained for entire reviews, to each snippet individually. This is approximately measuring the contribution of the snippet to the entire review score. Units containing trigger words with large corresponding lambda values are ranked higher. The intuition is that these units were decisive for the classifier for its predicted rating, e.g. [8], so they should be considered as useful evidence to present to humans. Examples are shown on Figure 7.

4 Categorizing information

In order to present a compact and comprehensive view of a restaurant, we should touch upon all relevant aspects of a restaurant with little redundancy. As we will see, simple statistical methods would normally fail to account for scarcely mentioned aspects, because reviews tend to be skewed, understandably, towards food comments. Therefore, we need to find some scheme which enforces diversity in the summary. Moreover, it is our hope that with aspect classification, we will be able in the future to build more complex services. Thus, we take a more ontological approach.

4.1 Labeling data

For restaurants, we have universally accepted axes:

1. Food: how the food tastes, and comments about specific items or selection.
2. Service: mostly politeness and timeliness of delivery.
3. Atmosphere: information about the venue – parking, cleanliness, loudness of music, etc.
4. Value: the quality of goods delivered by price.

We selected 23000 snippets and labeled them according to the four above-mentioned classes, as well as other, called X (e.g. “We were there on a Tuesday night”), or more than one class, called Z (“Food was spoiled by the atmosphere”). For a class to be classified in its category, it had to be useful to present as a snippet for a restaurant (e.g. “Panna cotta is a desert” doesn’t qualify, but “Crab cakes were cooked to perfection” does) otherwise it would be classified as “other” class (X). The distribution of labels is shown on Table 2.

Category	Percentage
Food (F)	33.98%
Service (S)	8.89%
Atmosphere (A)	11.84%
Value (V)	3.91%
Other (X)	33.78%
Multiple (Z)	7.56%

Table 2: Prior category distribution: obviously, there is a strong bias for describing food.

For experiments, we kept 1000 units for development and 1000 for evaluation. The inter-annotator agreement rate was almost 100%, the same as intra-annotator agreement.

4.2 A log-linear classifier

To categorize reviews, we used a log-linear classifier with unigram trigger features similar to the one used for numerical rating prediction. We had 32k features. The classifier results are shown on Figure 8. The error rate in a six-way classification is 27%. The prior error rate (by choosing food unconditionally) is 63%. In addition, thanks to the dual guideline in labeling, the classifer may also be used to determine whether a snippet is useful. The ROC curves are shown in Figure 9. For specific categories, since they are many reviews to choose from, and a single output candidate, we are more interested in the precision. Figure 10 shows the precision as the posterior threshold is increased.

4.3 Probabilistic Latent Semantic Analysis

Probabilistic Latent Semantic Analysis (PLSA) [9] is a technique which analyzes the relationship between documents and words using a latent factorization. Thus, PLSA could be used to derive a clustering which could in turn be used for gaining more insight into the task or for improving classification performance.

In the context of our task, PLSA builds a mixture model for a review snippet as:

$$p(s, w) = \sum_{z=1}^{M} p(w|z)p(z|s), \tag{3}$$

where s is the review snippet, w are the words, and z the hidden variables. The PLSA model is trained by maximizing the likelihood of the training data. Model components $p(\cdot|z)$ are thought to explain aspects of the review snippets, such as a topic.

To measure the information provided by PLSA, we compute the mutual information between the latent space representation of the review snippet and its category. The latent space representation can be encoded in 2 ways: 1) by clustering the latent space vectors or 2) by using the most likely z.

The entropy of the categories was 2.20 bits. We used 100 dimensions for PLSA. The mutual information between PLSA representation 1) and the category was 0.21 bits. The mutual information between PLSA representation 2) and the category was also 0.21 bits. This shows that PLSA provides a small amount of information about the categories. We therefore added features representing the $p(z|s)$ values to the log-linear classifier. The cross entropy of the log-linear classifier on the evaluation data was 3.43 bits. Adding the PLSA decomposition as additional 100 features reduces the cross-entropy by 0.17 bits, which is to say the classifier learned from PLSA additional information comparable to the use of PLSA alone. However, the category classification error rate only decreased from 27.3% to 26.5%.

Another aspect of PLSA which were interested in was the ability to automatically discover the categories in a completely unsupervised manner. After looking at the PLSA dimensions $p(\cdot|z)$, we were unable to assign interpretable knowledge to these components. On Figure 11, we show the words with the highest likelihood ratio against the unigram background model in some example dimensions. Thus, we were unable to significantly improve the category classification using PLSA nor were we able to discover the categories through PLSA.

4.4 Final condensation

In order to preserve a balanced view touching upon all possible aspects of a restaurants, we used both the relevance with polarity, and category information. From the steps above, we enumerate the most extreme snippets in food, then proceed to service, atmosphere, and price. An example summary is given in Figure 12.

5 Mean Opinion Score (MOS)

The system was evaluated as a snippet summarization system by evaluating user preferences. Ten test subjects were each asked to examine approximately 90 restaurants, for a total of 867 restaurants. For each restaurant, we presented the name of the restaurant along with its city location and global score, and then side-by-side our summarized snippets versus a random, non-overlapping selection of snippets, with the same number of snippets. Sides were randomized. We randomly selected restaurants to present over the set of restaurants which had more than 6 snippets available.

Subjects were asked to tell whether they preferred, or much preferred, one set of snippets over the other. Results (Table 3) were found to be significant with the Fisher sign test and the Wilcoxon signed rank test. Test subjects seemed to indicate a marked preference for snippets generated with our proposed selection technique.

Preference	Percentage
Proposed much better	27%
Proposed better	24%
Neither is better	22%
Random better	13%
Random much better	14%

Table 3: User preferences: 10 subjects compared our proposed approach to random snippet selection.

6 Conclusion

We have presented a review summarization system for the restaurant domain. The system summarizes a collection of user reviews using category and sentiment axes. We evaluated our system using a large-scale database collected from the Internet. The system was employed to identify areas of importance within a review to produce a summary for a restaurant. In a user study, subjects preferred our snippet selection over a random snippet selection by a statistically significant margin.

There are many open issues which deserve further study. The most prominent problem resides in snippet segmentation: there is no constraint about well-formedness, and indeed many snippets appear ungrammatical or outlandish when taken out of context. The question remains open as to how this approach may be extended to other types of products (services or consumer products). In the future, it may be possible to provide additional features such as faceted restaurant search or ability to drill down on details about a restaurant.

References

[1] S. Morinaga, K. Ya Yamanishi, K. Tateishi, and T. Fukushima, “Mining Product Reputations on the Web,” in *KDD*, 2002.

[2] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *SIGKDD*, 2004.

[3] K. Dave, S. Lawrence, and D. Pennock, "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews," in *WWW*, 2003.

[4] B. Pang, L. Lee, , and S. Vaithyanathan, "Thumbs Up? Sentiment Classification Using Machine Learning Techniques," in *Proceedings of EMNLP*, 2002.

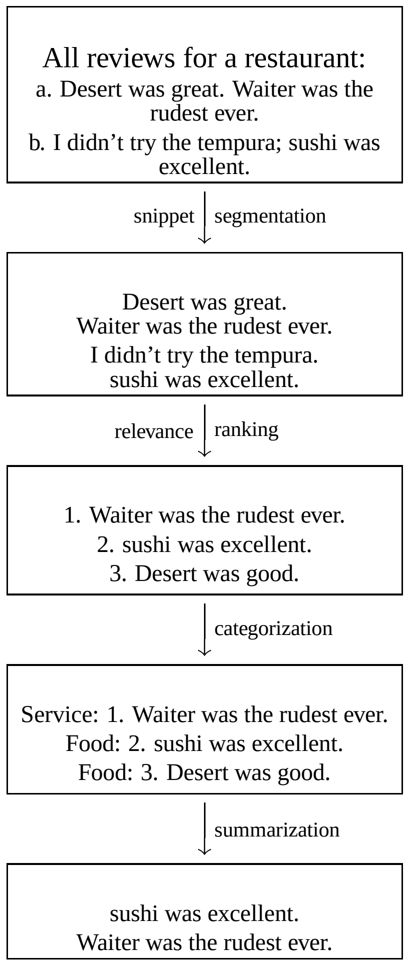
[5] S. Chen and R. Rosenfeld, "A survey of smoothing techniques for ME models," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 2, pp. 37–50, January 2000.

[6] A. Berger, S. Della Pietra, and V. Della Pietra, "A maximum entropy approach to natural language processing," *Computational Linguistics*, vol. 22, no. 1, pp. 39–71, 1996.

[7] P. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews," in *ACL*, 2002.

[8] V. Hatzivassiloglou and K. McKeown, "Predicting the Semantic Orientation of Adjectives," in *Proceedings of* 35^{th} *ACL/*8^{th} *EACL*, 1997.

[9] T. Hofmann, "Probabilistic latent semantic analysis," in *Proc. of Uncertainty in Artificial Intelligence, UAI'99*, Stockholm, 1999.



All reviews for a restaurant:
a. Desert was great. Waiter was the rudest ever.
b. I didn’t try the tempura; sushi was excellent.

Desert was great.
Waiter was the rudest ever.
I didn’t try the tempura.
sushi was excellent.

1. Waiter was the rudest ever.
2. sushi was excellent.
3. Desert was good.

Service: 1. Waiter was the rudest ever.
Food: 2. sushi was excellent.
Food: 3. Desert was good.

sushi was excellent.
Waiter was the rudest ever.

Figure 1: System overview: snippets are ranked and categorized before summarization.

It is a tiny Thai cafe on ninth avenue. The food here is very tasty Thai food. It looks slick, modern, and very attractive cafe. I'm a real fan of this Tiny... Your waiter staff ....is so friendly. I'm a fan.

Hostess: Ajax changed our reservation from 7pm to 8pm no reason given compromised on 7:45pm. Why? Truth: as stated by our hostess on our departure "we like to be out here by 9pm on Tuesdays".

The room, service and ambience truly speak of a bygone elegance and capture the spirit of Old New York. I went to see Eartha Kitt and was as impressed by the service as I was by her amazing performance. The food was basic, America fare - very well-prepared but nothing too exciting. And it was extremely expensive, but worth every penny. I would strongly recommend Cafe Carlyle as THE BEST special occasion restaurant in the city.

Figure 2: Sample reviews: reviews are typically compact.

Almost everything was good but deserts which tasted awful. Otherwise, service was excellent.

Figure 3: Issues with bag of words: good and bad impressions (underlined) mixed. A finer segmentation would be appropriate.

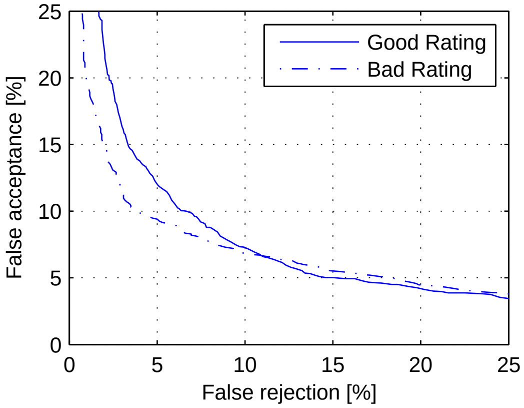
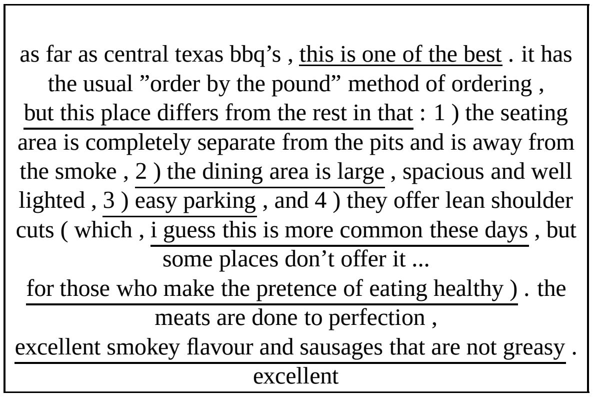


Figure 4: ROC curve for good and bad ratings. The Equal Error Rate (EER) is between 6-8%.



as far as central texas bbq's , this is one of the best . it has the usual ”order by the pound” method of ordering , but this place differs from the rest in that : 1) the seating area is completely separate from the pits and is away from the smoke , 2) the dining area is large , spacious and well lighted , 3) easy parking , and 4) they offer lean shoulder cuts (which , i guess this is more common these days , but some places don't offer it ... for those who make the pretence of eating healthy) . the meats are done to perfection , excellent smokey flavour and sausages that are not greasy . excellent

Figure 5: Example snippet decomposition: snippets are in alternating underline font. Punctuation marks delimit snippets.

worst pizza
was excited
worst restaurant
courteous .
worst food
worst dining
never recommend
after being
waiter ,
expensive .
. oh
with many
terrible food

Figure 6: Features with highest magnitude of associated λ weight.

Score	Snippet
0.02	i would go during a weekday when it’s not so crowded and you have the place to yourselves
-0.26	what kind of ice cream shop doesb’t offer samples
0.31	they crank out some creamy and delicious ice cream

Figure 7: Example of snippet relevance scores. Scores near 0 indicate blandness, positive scores a good opinion, negative scores a bad opinion.

[%]	F	A	S	V	X	Z
F	**84.9**	0.3	1.2	0.6	11.1	1.9
A	12.8	**34.0**	4.3	0.0	47.9	1.1
S	11.5	2.9	**62.5**	0.0	16.4	6.7
V	19.2	0.00	3.9	**23.1**	50.0	3.9
X	16.4	1.1	1.4	0.4	**80.7**	0.0
Z	29.6	4.6	2.3	0.0	4.6	**59.1**

Figure 8: Classifier results for categorization. Lines correspond to labeled category, columns are the resulting decoding. Overall error: 27%.

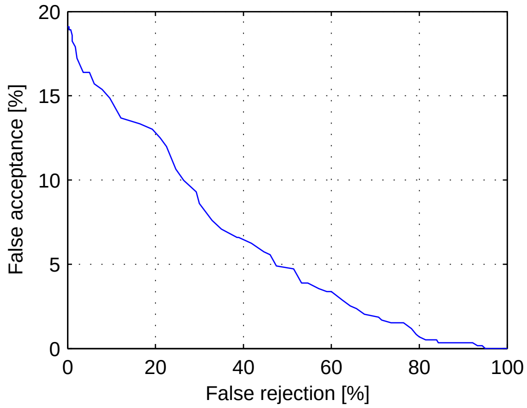


Figure 9: Relevance rejection characteristics for class X.

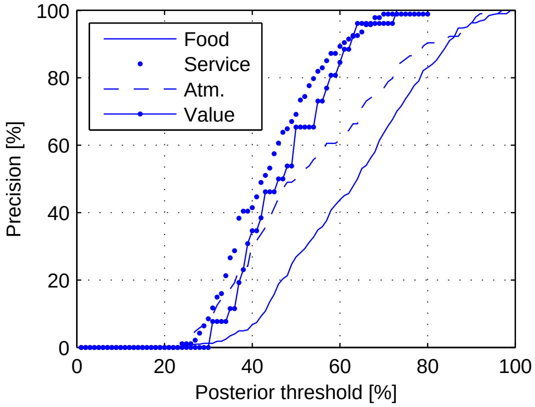


Figure 10: Precision with respect to posterior threshold.

Dim 1	Dim 2	Dim 3
everything	were	bring
although	both	sit
busy	cold	put
crowded	appetizers	watch
inside	course	able
packed	meals	please
quickly	main	pick
rich	roll	help

Figure 11: Example PLSA dimensions: a dimension would ideally map to a category.

- the food is outstanding and drinks are well made.
- the fried chicken is divine and the desserts are great too.
- service is a little moody at times.
- the wait staff was excellent and catered to us and the baby.
- the service is always excellent.
- the sous chef is great and even brought my table a complementary appetizer one night.
- lively atmosphere with a fantastic menu.
- be prepared for the extremely high prices.
- a little steep price wise.

Figure 12: Example of condensed summary produced by our system.