Word Salad: Relating Food Prices and Descriptions

Victor Chahuneau **Kevin Gimpel**
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{vchahune,kgimpel}@cs.cmu.edu

Bryan R. Routledge
Tepper School of Business
Carnegie Mellon University
Pittsburgh, PA 15213, USA
routledge@cmu.edu

Lily Scherlis
Phillips Academy
Andover, MA 01810, USA
lily.scherlis@gmail.com

Noah A. Smith
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
nasmith@cs.cmu.edu

Abstract

We investigate the use of language in food writing, specifically on restaurant menus and in customer reviews. Our approach is to build predictive models of concrete external variables, such as restaurant menu **prices**. We make use of a dataset of menus and customer reviews for thousands of restaurants in several U.S. cities. By focusing on prediction tasks and doing our analysis at scale, our methodology allows quantitative, objective measurements of the words and phrases used to describe food in restaurants. We also explore interactions in language use between menu prices and **sentiment** as expressed in user reviews.

1 Introduction

What words might a menu writer use to justify the high price of a steak? How does describing an item as *chargrilled* vs. *charbroiled* affect its price? When a customer writes an unfavorable review of a restaurant, how is her word choice affected by the restaurant's prices? In this paper, we explore questions like these that relate restaurant menus, prices, and customer sentiment. Our goal is to understand how language is used in the food domain, and we direct our investigation using external variables such as restaurant menu **prices**.

We build on a thread of NLP research that seeks linguistic understanding by predicting real-world quantities from text data. Recent examples include prediction of stock volatility (Kogan et al., 2009) and movie revenues (Joshi et al., 2010). There, prediction tasks were used for quantitative evaluation and objective model comparison, while analysis of learned models gave insight about the social process behind the data.

We echo this pattern here as we turn our attention to language use on restaurant menus and in user restaurant reviews. We use data from a large corpus of restaurant menus and reviews crawled from the web and formulate several prediction tasks. In addition to predicting menu prices, we also consider predicting **sentiment** along with price.

The relationship between language and sentiment is an active area of investigation (Pang and Lee, 2008). Much of this research has focused on customer-written reviews of goods and services, and perspectives have been gained on how sentiment is expressed in this type of informal text. In addition to sentiment, however, other variables are reflected in a reviewer's choice of words, such as the price of the item under consideration. In this paper, we take a step toward joint modeling of multiple variables in review text, exploring connections between price and sentiment in restaurant reviews.

Hence this paper contributes an exploratory data

Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 1357–1367, Jeju Island, Korea, 12–14 July 2012.

analysis of language used to describe food (by its purveyors and by its consumers). While our primary goal is to understand the language used in our corpus, our findings bear relevance to economics and hospitality research as well. This paper is a step on the way to the eventual goal of using linguistic analysis to understand social phenomena like sales and consumption.

2 Related Work

There are several areas of related work scattered throughout linguistics, NLP, hospitality research, and economics.

Freedman and Jurafsky (2011) studied the use of language in food advertising, specifically the words on potato chip bags. They argued that, due to the ubiquity of food writing across cultures, ethnic groups, and social classes, studying the use of language for describing food can provide perspective on how different socioeconomic groups self-identify using language and how they are linguistically targeted. In particular, they showed that price affects how "authenticity" is realized in marketing language, a point we return to in §5. This is an example of how price can affect how an underlying variable is expressed in language. Among other explorations in this paper, we consider how price interacts with expression of sentiment in user reviews of restaurants.

As mentioned above, our work is related to research in predicting real-world quantities using text data (Koppel and Shtrimberg, 2006; Ghose et al., 2007; Lerman et al., 2008; Kogan et al., 2009; Joshi et al., 2010; Eisenstein et al., 2010; Eisenstein et al., 2011; Yogatama et al., 2011). Like much of this prior work, we aim to learn how language is used in a specific context while building models that achieve competitive performance on a quantitative prediction task.

Along these lines, there is recent interest in exploring the relationship between product sales and user-generated text, particularly online product reviews. For example, Ghose and Ipeirotis (2011) studied the sales impact of particular properties of review text, such as readability, the presence of spelling errors, and the balance between subjective and objective statements. Archak et al. (2011) had a similar goal but decomposed user reviews into parts describing particular aspects of the product being reviewed (Hu and Liu, 2004). Our paper differs from price modeling based on product reviews in several ways. We consider a large set of weakly-related products instead of a homogeneous selection of a few products, and the reviews in our dataset are not product-centered but rather describe the overall experience of visiting a restaurant. Consequently, menu items are not always mentioned in reviews and rarely appear with their exact names. This makes it difficult to directly use review features in a pricing model for individual menu items.

Menu planning and pricing has been studied for many years by the culinary and hospitality research community (Kasavana and Smith, 1982; Kelly et al., 1994), often including recommendations for writing menu item descriptions (Miller and Pavesic, 1996; McVety et al., 2008). Their guidelines frequently include example menus from successful restaurants, but typically do not use large corpora of menus or automated analysis, as we do here. Other work focused more specifically on particular aspects of the language used on menus, such as the study by Zwicky and Zwicky (1980), who made linguistic observations through manual analysis of a corpus of 200 menus.

Relatedly, Wansink et al. (2001; 2005) showed that the way that menu items are described affects customers' perceptions and purchasing behavior. When menu items are described evocatively, customers choose them more often and report higher satisfaction with quality and value, as compared to when they are given the same items described with conventional names. Wansink et al. did not use a corpus, but rather conducted a small-scale experiment in a working cafeteria with customers and collected surveys to analyze consumer reaction. While our goals are related, our experimental approach is different, as we use automated analysis of thousands of restaurant menus and rely on a set of one million reviews as a surrogate for observing customer behavior.

Finally, the connection between products and prices is also a central issue in economics. However, the stunning heterogeneity in products makes empirical work challenging. For example, there are over 50,000 menu items in New York that include

City	# Restaurants			# Menu Items			# Reviews		
	train	dev.	test	train	dev.	test	train	dev.	test
Boston	930	107	113	63,422	8,426	8,409	80,309	10,976	11,511
Chicago	804	98	100	51,480	6,633	6,939	73,251	9,582	10,965
Los Angeles	624	80	68	17,980	2,938	1,592	75,455	13,227	5,716
New York	3,965	473	499	365,518	42,315	45,728	326,801	35,529	37,795
Philadelphia	1,015	129	117	83,818	11,777	9,295	52,275	7,347	5,790
San Francisco	1,908	255	234	103,954	12,871	12,510	499,984	59,378	67,010
Washington, D.C.	773	110	121	47,188	5,957	7,224	71,179	11,852	14,129
Total	10,019	1,252	1,252	733,360	90,917	91,697	1,179,254	147,891	152,916

Table 1: Dataset statistics.

the word *chicken*. What is the price of chicken? This is an important practical and daunting matter when measuring inflation (e.g., Consumer Price Index is measured with a precisely-defined basket of goods). Price dispersion across goods and the variation of the goods is an important area of industrial organization economic theory. For example, economists are interested in models of search, add-on pricing, and obfuscation (Baye et al., 2006; Ellison, 2005).

3 Data

We crawled Allmenus.com (`www.allmenus.com`) to gather menus for restaurants in seven U.S. cities: Boston, Chicago, Los Angeles, New York, Philadelphia, San Francisco, and Washington, D.C. Each menu includes a list of item names with optional text descriptions and prices. Most Allmenus restaurant pages contain a link to the corresponding page on Yelp (`www.yelp.com`) with metadata and user reviews for the restaurant, which we also collected.

The metadata consist of many fields for each restaurant, which can be divided into three categories: location (city, neighborhood, transit stop), services available (take-out, delivery, wifi, parking, etc.), and ambience (good for groups, noise level, attire, etc.). Also, the category of food and a price range ($ to $$$$, indicating the price of a typical meal at the restaurant) are indicated. The user reviews include a star rating on a scale of 1 to 5.

The distribution of prices of individual menu items is highly skewed, with a mean of $9.22 but a median of $6.95. On average, a restaurant has 73 items on its menu with a median price of $8.69 and 119 Yelp reviews with a median rating of 3.55 stars. The star rating and price range distributions are shown in Figure 1.

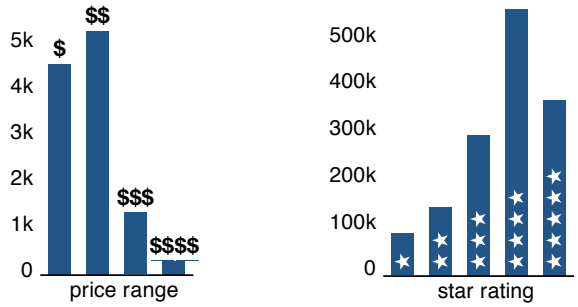

Figure 1: Frequency distributions of restaurant price ranges (left) and review ratings (right).

The set of restaurants was randomly split into three parts (80% for training, 10% for development, 10% for evaluation), independently for each city. The sizes of the splits and the full set of dataset statistics are provided in Table 1.

4 Predictive Tasks

We consider several prediction tasks using the dataset just described. These include predicting **individual menu item prices** (§5), predicting the **price range** for each restaurant (§6), and finally jointly predicting **median price and sentiment** for each restaurant (§7). To do this, we use two types of models: linear regression (§5 and §6) and logistic regression (§7), both with ℓ_1 regularization when sparsity is desirable. We tune the regularization coefficient by choosing the value that minimizes development set loss (mean squared error and log loss, respectively).

For evaluation, we use mean absolute error (MAE) and mean relative error (MRE). Given a dataset $\langle \boldsymbol{x}_i, y_i \rangle_{i=1}^{N}$ with inputs $\boldsymbol{x}_i$ and outputs y_i, and

denoting predicted outputs by $\hat{y}_i$, these are defined as follows:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

$$\text{MRE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

In practice, since we model log-prices but evaluate on real prices, the final prediction is often a non-linear transformation of the output of the linear classifier of weight vector $\boldsymbol{w}$, which we denote by: $\hat{y}_i = f(\boldsymbol{w}^\top \boldsymbol{x}_i)$.

We also frequently report the total number of features available in the training set for each model (nf) as well as the number of non-zero feature weights following learning (nnz).

5 Menu Item Price Prediction

We first consider the problem of predicting the **price** of each item on a menu. In this case, every instance $\boldsymbol{x}_i$ corresponds to a single item in the menu parametrized by the features detailed below and y_i is the item's price. In this section, our models always use the logarithm of the price[1] as output values and therefore: $\hat{y}_i = e^{\boldsymbol{w}^\top \boldsymbol{x}_i}$.

Baselines We evaluate several baselines which make independent predictions for each distinct item name. The first two predict the mean or the median of the prices in the training set for a given item name, and use the overall price mean or median when a name is missing in the training set. The third baseline is an ℓ_1-regularized linear regression model trained with a single binary feature for each item name in the training data. These baselines are shown as the first three rows in Table 2.

We note that there is a wide variation of menu item names in the dataset, with more than 400,000 distinct names. Although we address this issue later by introducing local text features, we also performed simple normalization of the item names for all of the baselines described above. To do this normalization, we first compiled a stop word list based on the most frequent words in the item names.[2] We removed stop words and then ordered the words in each item name lexicographically, in order to collapse together items such as *coffee black* and *black coffee*. This normalization reduced the unique item name count by 40%, strengthening the baselines.

5.1 Features

We use ℓ_1-regularized linear regression for feature-rich models. We now introduce several sets of features that we add to the normalized item names:[3]

I. METADATA: Binary features for each restaurant metadata field mentioned above, excluding price range. A separate binary feature is included for each unique ⟨field, value⟩ tuple.
II. MENUNAMES: n-grams in menu item names. We used binary features for unique unigrams, bigrams, and trigrams. Here, stop words were retained as they can be informative (e.g., *with* and *large* correlate with price).
III. MENUDESC: n-grams in menu item descriptions, as in MENUNAMES.

Review Features In addition to these features, we consider leveraging the large amount of text present in user reviews to improve predictions. We attempted to find mentions of menu items in the reviews and to include features extracted from the surrounding words in the model. Perfect item mentions being relatively rare, we consider inexact matches weighted by a coefficient measuring the degree of resemblance: we used the Dice similarity between the set of words in the sentence and in the item name. We then extracted n-gram features from this sentence, and tried several ways to use them for price prediction.

Given a review sentence, one option is to add the corresponding features to every item matching this sentence, with a value equal to the similarity coefficient. Another option is to select the best matching item and use the same real-valued features but only for this single item. Binary feature values can be used instead of the real-valued similarity coefficient. We also experimented with the use of part-of-speech tags in order to restrict our features to adjective and adverb n-grams instead of the full vocabulary. All of

[1] The price distribution is more symmetric in the log domain.
[2] This list can be found in the supplementary material.
[3] The normalized item names are present as binary features in all of our regression models

	MAE	MRE	nf	nnz
Predict mean	3.70	43.32	n/a	n/a
Predict median	3.67	43.93	n/a	n/a
Regression	3.66	45.64	267,945	240,139
METADATA	3.55	43.11	268,450	258,828
$\widehat{\text{PR}}$	3.47	43.11	267,946	205,176
MENUNAMES	3.23	38.33	896,631	230,840
+ MENUDESC	3.19	36.23	1,981,787	151,785
+ $\widehat{\text{PR}}$	3.08	34.51	1,981,788	140,954
+ METADATA	3.08	34.97	1,982,363	148,774
+ MENTIONS	3.06	34.37	4,959,488	458,462

Table 2: Results for menu item price prediction. MAE = mean absolute error ($), MRE = mean relative error (%), nf = total number of features, nnz = number of features with non-zero weight.

these attempts yielded negative or only slightly positive results, of which we include only one example in our experiments: the MENTIONS feature set consists of n-grams for the best matching item with the Dice coefficient as the feature value.

We also tried to incorporate the reviews by using them in aggregate via predictions from a separate model; we found this approach to work better than the methods described above which all use features from the reviews directly in the regression model. In particular, we use the review features in a separate model that we will describe below (§6) to predict the price range of each restaurant. The model uses unigrams, bigrams, and trigrams extracted from the reviews. We use the estimated price range (which we denote $\widehat{\text{PR}}$) as a single additional real-valued feature for individual item price prediction.

5.2 Results

Our results are shown in Table 2. We achieve a final reduction of 50 cents in MAE and nearly 10% in MRE compared with the baselines. Using menu name features (MENUNAMES) brings the bulk of the improvement, though menu description features (MENUDESC) and the remaining features also lead to small gains. Interestingly, as the MENUDESC and $\widehat{\text{PR}}$ features are added to the model, the regularization favors more general features by selecting fewer and fewer non-zero weights.

While METADATA features improve over the baselines when used alone, they do not lead to improved performance over the MENU* + $\widehat{\text{PR}}$ features, suggesting that the text features may be able to substitute for the information in the metadata, at least for prediction of individual item prices.

The MENTIONS features resulted in a small improvement in MAE and MRE, but at the cost of expanding the model size significantly. A look at the learned feature weights reveals that most of the selected features seem more coincidental than generic (*rachel's*, highly negative) when not totally unintuitive (*those suicide*, highest positive). This suggests that our method of extracting features from mentions is being hampered by noise. We suspect that these features could be more effective with a better method of linking menu items to mentions in review text.

5.3 Analysis

We also inspected the feature weights of our learned models. By comparing the weights of related features, we can see the relative differences in terms of contribution to menu item prices. Table 3 shows example feature weights, manually arranged into several categories (taken from the model with MENUNAMES + MENUDESC + $\widehat{\text{PR}}$ + METADATA).

Table 3(a) shows selected features for the "ambience" field in the Yelp restaurant metadata and pane (b) lists some unigrams related to cooking methods. Pane (c) shows feature weights for n-grams often used to market menu items; we see larger weights for words targeting those who want to eat organically- or locally-grown food (*farmhouse*, *heirloom*, *wild*, and *hormone*), compared to those looking for comfort food (*old time favorite*, *traditional*, *real*, and *fashioned*). This is related to observations made by Freedman and Jurafsky (2011) that cheaper food is marketed by appealing to tradition and historicity, with more expensive food described in terms of naturalness, quality of ingredients, and the preparation process (e.g., *hand picked*, *wild caught*, etc.). Relatedly, in pane (e) we see that *real* mashed potatoes are expected to be cheaper than those described as *creamy* or *smooth*.

Pane (d) shows feature weights for trigrams containing units of chicken; we can see an ordering in terms of size (*bits* < *cubes* < *strips* < *cuts*) as well as the price increase associated with the use of the word *morsels* in place of less refined units. We also see a difference between *pieces* and *pcs*, with the latter being frequently used to refer to entire cuts of

(a) METADATA: ambience	
dive-y	-0.015
intimate	-0.013
trendy	-0.012
casual	-0.005
romantic	-0.004
classy	-7e-6
touristy	0.058
upscale	0.099

(b) MENUDESC: cooking	
panfried	-0.094
chargrilled	-0.029
cooked	-0.012
boiled	-0.006
fried	-0.005
steamed	0.011
charbroiled	0.015
grilled	0.022
simmered	0.025
roasted	0.034
sauteed	0.034
broiled	0.053
seared	0.066
braised	0.068
stirfried	0.071
flamebroiled	0.106

(c) MENUDESC: descriptors	
old time favorite	-0.112
fashioned	-0.034
line caught	-0.028
all natural	-0.028
traditional	-0.009
natural	3e-4
classic	0.002
free range	0.004
real	0.004
fresh	0.006
homemade	0.010
authentic	0.012
organic	0.020
specialty	0.025
special	0.033
locally	0.037
natural grass fed	0.038
artisanal	0.064
raised	0.066
heirloom	0.083
wild	0.084
hormone	0.085
farmed	0.099
hand picked	0.101
wild caught	0.116
farmhouse	0.133

(d) MENUDESC: _ = "of chicken"	
slices _	-0.102
bits _	-0.032
cubes _	-0.030
pieces _	-0.024
strips _	-0.001
chunks _	0.015
morsels _	0.025
pcs _	0.040
cuts _	0.042

(e) MENUDESC: _ = "potatoes"	
real mashed _	-0.028
mashed _	-0.005
creamy mashed _	-5e-9
smashed _	0.018
smooth mashed _	0.129

(f) MENUDESC: _ = "potato"	
mash _	-0.022
mashed _	-0.019

(g) MENUDESC: "crisp" vs. "crispy"	
crisp	-0.022
crispy	-0.011
crispy bacon	0.008
crisp bacon	0.033

(h) MENUDESC: "roast" vs. "roasted"	
roasted	0.034
roast	0.040
roasted potatoes	0.026
roast potatoes	0.110
roasted chicken	-0.041
roast chicken	-0.012
roasted salmon	0.091
roast salmon	0.151
roast pork	-0.038
roasted pork	0.055
roasted tomato	0.010
roast tomato	0.026

Table 3: Selected features from model for menu item price prediction. See text for details.

chicken (e.g., wings, thighs, etc.) and the former more often used as a synonym for *chunks*.

Panes (f), (g), and (h) reveal price differences due to slight variations in word form. We find that, even though *crispy* has a higher weight than *crisp*, *crisp bacon* is more expensive than *crispy bacon*. We also find that food items prefixed with *roast* lead to more expensive prices than the similar *roasted*, except in the case of *pork*, though here the different forms may be evoking two different preparation styles.

Also of note is the slight difference between the nonstandard *mash potato* and *mashed potato*. We observed lower weights with other nonstandard spellings, notably *portobella* having lower weight than each of the more common spellings *portabella*, *portobello*, and *portabello*.

6 Restaurant Price Range Prediction

In addition to predicting the prices of individual menu items, we also considered the task of predicting the **price range** listed for each restaurant on its Yelp page. The values for this field are integers from 1 to 4 and indicate the price of a typical meal from the restaurant.

For this task, we again train an ℓ_1-regularized linear regression model with integral price ranges as the true output values y_i. Each input $\boldsymbol{x}_i$ corresponds to the feature vector for an entire restaurant. For evaluation, we round the predicted values to the nearest integer: $\hat{y}_i = \text{ROUND}(\boldsymbol{w}^\top \boldsymbol{x}_i)$ and report the corresponding mean absolute error and accuracy.

We compared this simple approach with an ordinal regression model (McCullagh, 1980) trained with the same ℓ_1 regularizer and noted very little improvement (77.32% vs. 77.15% accuracy for METADATA). Therefore, we only report in this section results for the linear regression model.

In addition to the feature sets used for individual menu item price prediction, we used features on reviews (REVIEWS). Specifically, we used binary features for unigrams, bigrams, and trigrams in the full set of reviews for each restaurant. A stopword list was derived from the training data.[4] Bigrams and trigrams were filtered if they ended with stopwords. Additionally, features occurring fewer than three times in the training set were discarded.

[4]This list is included in the supplementary material.

Features	MAE	Acc.	nf	nnz
Predict mode	0.5421	48.22	n/a	n/a
MENU*	0.3875	66.29	1,910,622	995
METADATA	0.2372	77.15	591	219
REVIEWS	0.2172	79.76	3,027,470	1,567
+METADATA	0.2111	80.36	3,027,943	1,376

Table 4: Results for restaurant price range prediction. MAE = mean absolute error, Acc = classification accuracy (%), nf = total number of features, nnz = number of features with non-zero weight.

6.1 Results

Our results for price range prediction are shown in Table 4. Predicting the most frequent price range gave us an accuracy of 48.22%. Performance improvements were obtained by separately adding menu (MENU*), metadata (METADATA), and review features (REVIEWS). Unlike individual item price prediction, the reviews were more helpful than the menu features for predicting overall price range. This is not surprising, since reviewers will often generally discuss price in their reviews. We combined metadata and review features to get our best accuracy, exceeding 80%.

We also wanted to perform an analysis of sentiment in the review text. To do this, we trained a logistic regression model predicting polarity for each review; we used the REVIEWS feature set, but this time considering each review as a single training instance. The polarity of a review was determined by whether or not its star rating was greater than the average rating across all reviews in the dataset (3.7 stars). We achieved an accuracy of 87% on the test data. We omit full details of these models because the polarity prediction task for user reviews is well-known in the sentiment analysis community and our model is not an innovation over prior work (Pang and Lee, 2008). However, our purpose in training the model was to use the learned weights for understanding the text in the reviews.

6.2 Interpreting Reviews

Given learned models for predicting a restaurant's price range from its set of reviews as well as polarity for each review, we can turn the process around and use the feature weights to analyze the review text. Restricting our attention to reviews of 50–60 words, Table 5 shows sample reviews from our test set that lead to various predictions of price range and sentiment.[5]

This technique can also be useful when trying to determine the "true" star rating for a review (if provided star ratings are noisy), or to show the most positive and most negative reviews for a product within a particular star rating. The 5-point scale is merely a coarse approximation to the reviewer's mental state; using fitted models can provide additional clues to decode the reviewer's sentiment.

We can also do a more fine-grained analysis of review text by noting the contribution to the price range prediction of each position in the text stream. This is straightforward because our features are simply n-grams of the review text. In Figure 2, we show the influence of each word in a review sentence on the predicted polarity (brown) and price range (yellow). The height of a bar at a given position is proportional to the sum of the feature weights for every unigram, bigram, and trigram containing the token at that position (there are at most 6 active n-grams at a position).

The first example shows the smooth shift in expressed sentiment from the beginning of the sentence to the end. The second sentence is a difficult example for sentiment analysis, since there are several positive words and phrases early but the sentiment is chiefly expressed in the final clause. Our model noted the steady positive sentiment early in the sentence but identified the crucial negation due to strong negative weight on bigrams *fresh but*, *left me*, and *me yearning*. In both examples, the yellow bars show that price estimates are reflected mainly through isolated mentions of offerings and amenities (*drinks*, *atmosphere*, *security*, *good service*).

7 Joint Prediction of Price and Sentiment

Although we observe no interesting correlation ($r = 0.06$) between median star rating and median item price in our dataset, this does not imply that senti-

[5]To choose the 9 reviews in the table, we took the reviews from our test set in the desired length range and computed predicted sentiment and price range for each; then we scaled the predicted price range so that its range matched that of predicted sentiment, and maximized various linear combinations of the two. This accounts for the four corners. The others were found by maximizing a linear combination of one (possibly negated) prediction minus the absolute value of the other.

	← cheap		expensive →
↑ ⊕	i love me a cheap vietnamese sandwich . mmm , pate . this place has the best ones i 've had in the city , and i conveniently live a few blocks away . the ladies behind the counter are always courteous and fast , and who can beat a $ 3 sandwich ?! crazy ass deli .	this place is tiny ! the pork buns are so tender and flavorful . i dream about these things . manila clams were awesome , not the biggest clam fan either , but i loved it . mmm 7 spice chips . i ca n't wait to go back !	amazing service and desserts . nice wine list and urban decor . i went with a girlfriend and we split an entree , appetizer and dessert and they happily brought us separate portions which were just the right size . the bread is awesome , too . definitely a bit of a splurge , but worth it in moderation .
	great place to get fast food that tastes good . paneer and chicken are both good . i would prefer to go thursday thru saturday night . thats when they have their good shift working . also it stays open late until 4 am on weekends . really enjoyable !	had some solid thai here for lunch last week . ordered the special of the day , a chicken curry . quick service and nice interior . only issue was , had a bit of a stomach ache afterwards ? prefer their sister restaurant , citizen thai and the monkey , in north beach .	weekday evening was quiet , not every table was filled . our waiter was amicable and friendly , which is always a plus . the coconut bread pudding was ok and very sweet . it 's definitely a dessert plate that can be shared with a glass of wine .
↓ ⊖	for some reason my friend wanted me to go here with him . it was a decent standard greasy slice of pizza . it was n't bad by any means , but it was nothing special at all . on the plus side , cheap and fast . so in summary : cheap , fast , greasy , average .	ugh ! the salt ! all 5 dishes we ordered were so unbearably salty , i 'd rather just have the msg . greasy , oily , salty - there is much better chinese food to be had in sf than here . i was very disappointed and wo n't be back .	downhill alert ... had a decent lunch at dragon well this week marred by pretty spotty service . our waiter just did n't have it together , forgetting to bring bowls for our split soup , our beverages , etc. . food was good but pretty pricey for what we got .

Table 5: Reviews from the test set deemed by our model to have particular values of sentiment and price.

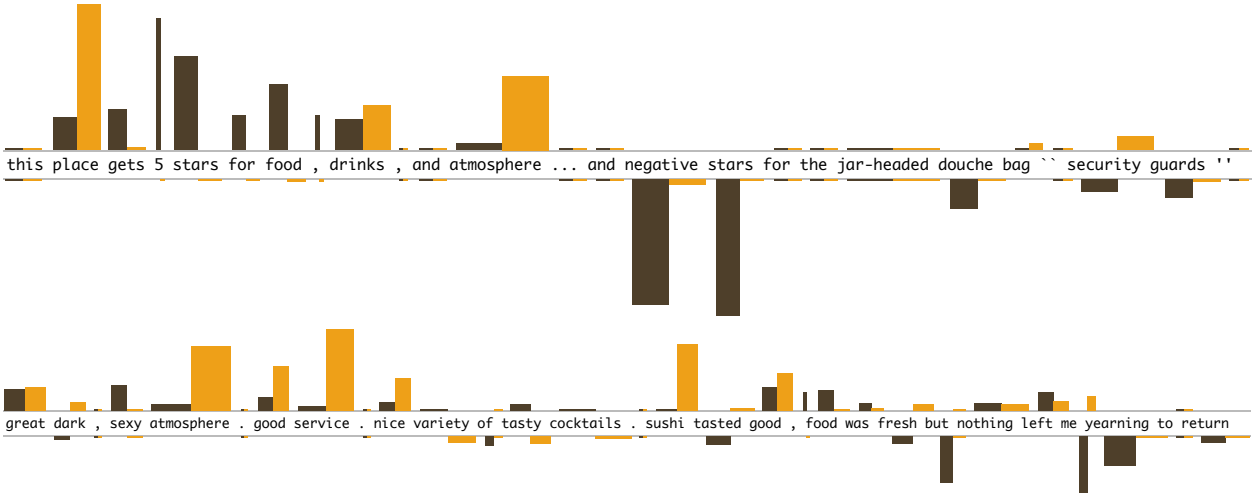

Figure 2: Local (position-level) sentiment (brown) and price (yellow) estimates for two sentences in the test corpus.

ment and price are independent of each other.[6] We try to capture this interaction by modeling at the same time review polarity and item price: we consider the task of jointly predicting **aggregate sentiment and price** for a restaurant.

For every restaurant in our dataset, we compute its median item price $\bar{p}$ and its median star rating $\bar{r}$. The average of these two values for the entire dataset ($8.69 and 3.55 stars) split the plane $(\bar{p}, \bar{r})$ in four sections: we assign each restaurant to one of these quadrants which we denote $\downarrow \ominus$, $\downarrow \oplus$, $\uparrow \ominus$ and $\uparrow \oplus$. This allows us to train a 4-class logistic regression model using the REVIEWS feature set for each restaurant. We achieve an accuracy of 65% on the test data, but we are mainly interested in interpreting the estimated feature weights.

7.1 Analysis

To visualize the top feature weights learned by the model, we have to map the four weight vectors learned by the model back to the underlying two-dimensional sentiment/price space. Therefore, we compute the following values:

$$\boldsymbol{w}_{\$} = (\boldsymbol{w}_{\uparrow\oplus} + \boldsymbol{w}_{\uparrow\ominus}) - (\boldsymbol{w}_{\downarrow\oplus} + \boldsymbol{w}_{\downarrow\ominus})$$
$$\boldsymbol{w}_{\bigstar} = (\boldsymbol{w}_{\uparrow\oplus} + \boldsymbol{w}_{\downarrow\oplus}) - (\boldsymbol{w}_{\uparrow\ominus} + \boldsymbol{w}_{\downarrow\ominus})$$

[6] Price and sentiment are both endogenous outcomes reflecting the characteristics of the restaurant. E.g., "better" restaurants can charge higher prices.

We then select for display the features which are the furthest from the origin ($\max w_{\$}^2 + w_{\bigstar}^2$) and represent the selected n-grams as points in the sentiment/price space to obtain Figure 3.

We notice that the spread of the sentiment values is larger, which suggests that reviews give stronger clues about consumer experience than about the cost of a typical meal. However, obvious price-related adjectives (*inexpensive* vs. *expensive*) appear in this limited selection, as well as certain phrases indicating both sentiment and price (*overpriced* vs. *very reasonable*). Other examples of note: *gem* is used in strongly-positive reviews of cheap restaurants; for expensive restaurants, reviewers use *highly recommended* or *amazing*. Also, phrases like *no flavor* and *manager* appear in negative reviews of more expensive restaurants, while *dirty* appears more often in negative reviews of cheaper restaurants.

8 Conclusion

We have explored linguistic relationships between food prices and customer sentiment through quantitative analysis of a large corpus of menus and reviews. We have also proposed visualization techniques to better understand what our models have learned and to see how they can be applied to new data. More broadly, this paper is an example of using extrinsic variables to drive model-building for linguistic data, and future work might explore richer extrinsic variables toward a goal of task-driven notions of semantics.

Acknowledgments

We thank Julie Baron, Ric Crabbe, David Garvett, Laura Gimpel, Chenxi Jiao, Elaine Lee, members of the ARK research group, and the anonymous reviewers for helpful comments that improved this paper. This research was supported in part by the NSF through CAREER grant IIS-1054319 and Sandia National Laboratories (fellowship to K. Gimpel).

References

N. Archak, A. Ghose, and P. G. Ipeirotis. 2011. Deriving the pricing power of product features by mining consumer reviews. *Management Science*, 57(8).

M. R. Baye, J. Morgan, and P. Scholten. 2006. Economics and information systems; handbooks in information systems. In T. Hendershott, editor, *Judgement under Uncertainty: Heuristics and Biases*. Elsevier, Amsterdam.

J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proc. of EMNLP*.

J. Eisenstein, N. A. Smith, and E. P. Xing. 2011. Discovering sociolinguistic associations with structured sparsity. In *Proc. of ACL*.

G. Ellison. 2005. A model of add-on pricing. *Quarterly Journal of Economics*, 120(2):585–637, May.

J. Freedman and D. Jurafsky. 2011. Authenticity in America: Class distinctions in potato chip advertising. *Gastronomica*, 11(4):46–54.

A. Ghose and P. G. Ipeirotis. 2011. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering*, 23(10).

A. Ghose, P. G. Ipeirotis, and A. Sundararajan. 2007. Opinion mining using econometrics: A case study on reputation systems. In *Proc. of ACL*.

M. Hu and B. Liu. 2004. Mining opinion features in customer reviews. In *Proc. of AAAI*.

M. Joshi, D. Das, K. Gimpel, and N. A. Smith. 2010. Movie reviews and revenues: An experiment in text regression. In *Proc. of NAACL*.

M. L. Kasavana and D. I. Smith. 1982. *Menu Engineering: A Practical Guide to Menu Analysis*. Hospitality Publications.

T. J. Kelly, N. M. Kiefer, and K. Burdett. 1994. A demand-based approach to menu pricing. *Cornell Hotel and Restaurant Administrative Quarterly*, 35(1).

S. Kogan, D. Levin, B. R. Routledge, J. Sagi, and N. A. Smith. 2009. Predicting risk from financial reports with regression. In *Proc. of NAACL*.

M. Koppel and I. Shtrimberg. 2006. Good news or bad news? let the market decide. *Computing Attitude and Affect in Text: Theory and Applications*.

K. Lerman, A. Gilder, M. Dredze, and F. Pereira. 2008. Reading the markets: Forecasting public opinion of political candidates by news analysis. In *Proc. of COLING*.

P. McCullagh. 1980. Regression models for ordinal data. *Journal of the royal statistical society. Series B (Methodological)*, pages 109–142.

P. J. McVety, B. J. Ware, and C. L. Ware. 2008. *Fundamentals of Menu Planning*. John Wiley & Sons.

J. E. Miller and D. V. Pavesic. 1996. *Menu: Pricing & Strategy*. Hospitality, Travel, and Tourism Series. John Wiley & Sons.

B. Pang and L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

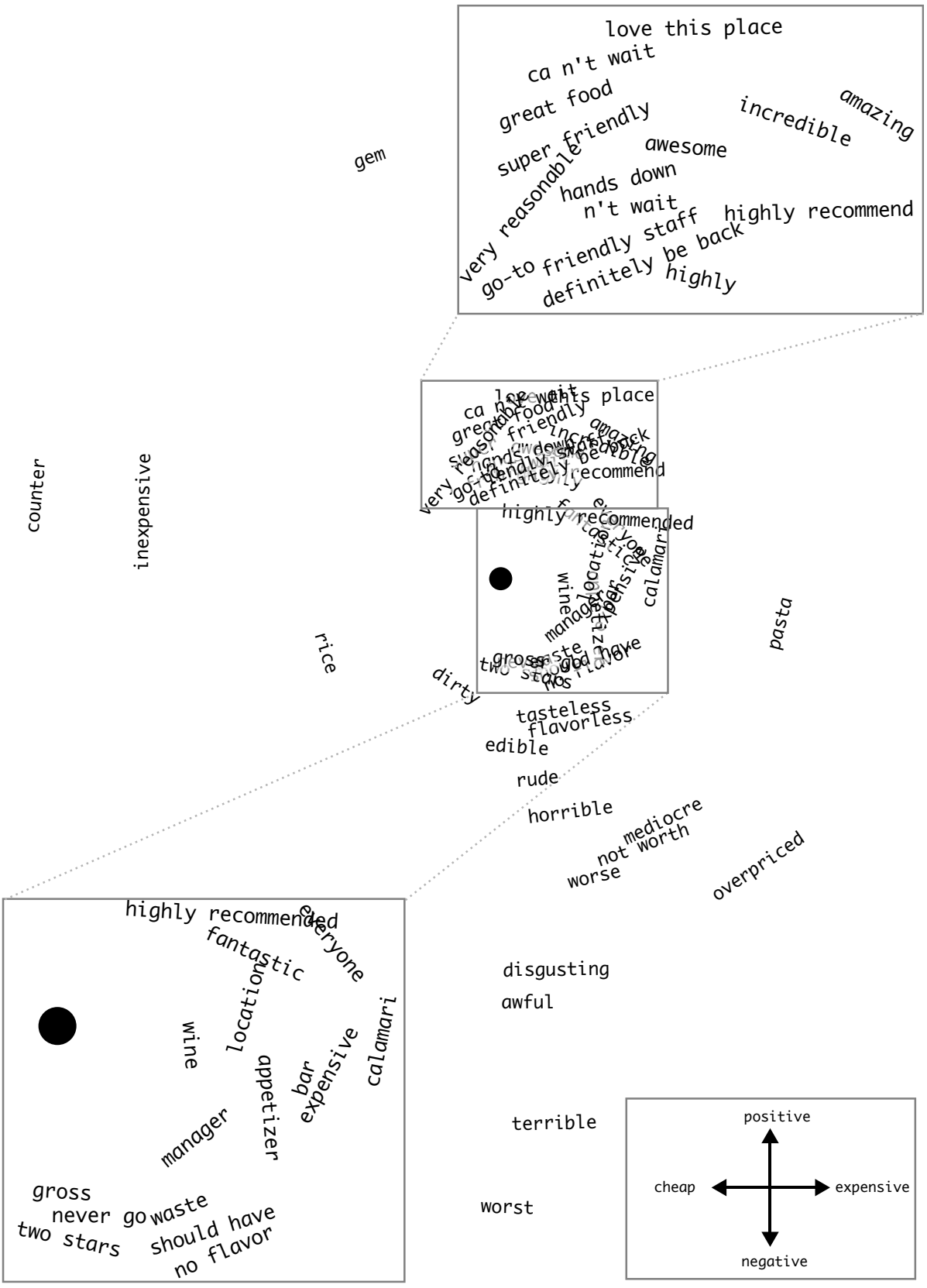

Figure 3: Top 50 features from joint prediction of price and sentiment. The black circle is the origin. See text for details on how the coordinates for each feature were computed. Insets show enlargements of dense areas of the graph.

B. Wansink, J. E. Painter, and K. van Ittersum. 2001. Descriptive menu labels' effect on sales. *Cornell Hotel and Restaurant Administrative Quarterly*, 42(6).

B. Wansink, K. van Ittersum, and J. E. Painter. 2005. How descriptive food names bias sensory perceptions in restaurants. *Food Quality and Preference*, 16(5).

D. Yogatama, M. Heilman, B. O'Connor, C. Dyer, B. R. Routledge, and N. A. Smith. 2011. Predicting a scientific community's response to an article. In *Proc. of EMNLP*.

A. D. Zwicky and A. M. Zwicky. 1980. America's national dish: The style of restaurant menus. *American Speech*, 55(2):83–92.