Word Salad: Relating Food Prices and Descriptions Supplementary Material

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

1 Trends in Predictions

We conjecture that systematic trends in errors can help detect patterns in restaurant prices. For example, if errors correlate with geography, we may identify regions with higher- or lower-than-expected prices. To test this for New York, we retrained our price range predictor using the entirety of New York as the test set, Philadelphia as the dev set, and the remaining cities as training data. We only used features from the reviews, omitting geographic metadata features. We then compare the predictions of the model to the real price range values in geographic context. More precisely, for every restaurant represented by its projected coordinates, we compute the relative error $\frac{\hat{y}-y}{y}$ at this point and smooth the resulting two-dimensional price distribution using a triangular kernel. We obtain the map of Manhattan shown in Figure 1, revealing trends in our predictions.

We first note the prevalence of blue in the plot; since we trained our models on cheaper cities and tested on Manhattan, we are systematically underpricing Manhattan restaurants. We see the most severe underprediction in Midtown East, an area known for its expensive cost of living. However, some areas differ from the overall trend; we observe moderate overprediction in Chinatown and parts of Greenwich Village and the East Village, areas which are generally known to be lower-priced than most of lower Manhattan.

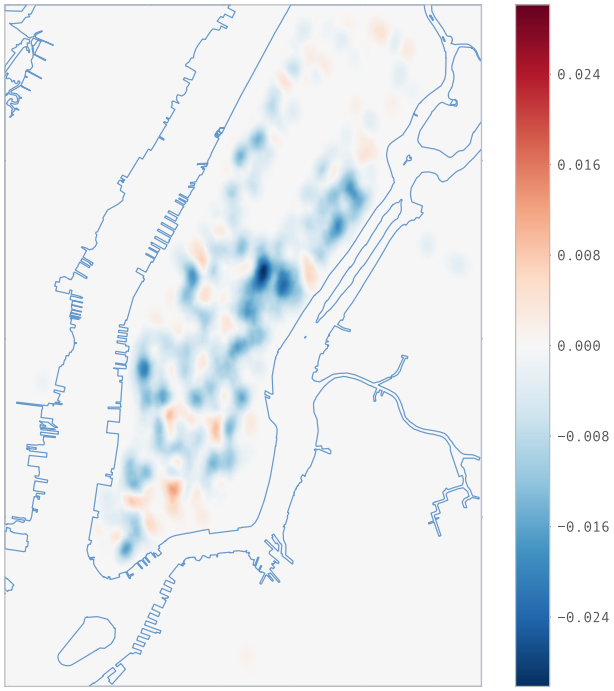

Figure 1: Smoothed distribution of relative differences between the value predicted from the reviews and the real price range for 2965 restaurants in Manhattan.

We clarify that the prediction is not made directly by using explicit mentions by users in their reviews (e.g., uses of key phrases such as "overpriced" and "great deal") but rather by considering predictions from a regression model using features aggregated across all reviews for the restaurant. While a simpler technique based on counts of key phrases may also be able to locate geographic trends such as these, our

approach can be used for any language and city since it does not require manual selection of key phrases. Our approach also allows us to inspect the degree to which our price estimate is wrong for a particular restaurant and to answer more general queries such as listing the most overpredicted or underpredicted restaurants in a particular neighborhood.

2 Stopword Lists Derived from Data

and, n', or, w, with, without
a, in, of, on, the
a, ai, al, alla, con, de, di, e, et, la, o, s, y
half, l, lb, oz, pack, pc, pcs, pint, pt, qt, quart
large, lg, med, medium, sm, small
all, here, i, it, my, that, there, they, this, we, you
a, an, and, at, for, in, of, on, or, so, the, to, with
'd, 'll, 's, 've, are, as, be, do, had, if, is, was, were, will

Table 1: Stopword lists used for feature extraction (top: menu item names, bottom: reviews).