CS839 Data Science Spring 2018

<Group 9>

Ya-Chun Yang, yang364@wisc.edu
Han Wang, hwang729@wisc.edu
Yanghui Kang, ykang38@wisc.edu

Stage 1 Report - Food Entity Extraction

INTRODUCTION

In this project stage, we extracted food name entity from customer reviews of a restaurant using supervised machine learning approach.

DATA SOURCE AND PREPROCESSING

We extracted 335 pieces of customer reviews for Marigold Kitchen, Madison, WI, from TripAdvisor. Food entity is defined as a word or a phrase that identifies food or a dish on the restaurant menu. Meal name words such as “breakfast”, “dinner” are considered as food. Examples of the food entity are listed in Table 1.

Table 1. Food Entity Examples

Document ID	Food Entity Example	Document ID	Food Entity Example
1	maple syrup	185	salmon frittata
18	sweet potato	225	omelettes
27	basil walnut pesto hash	247	scrambled eggs
117	breakfast sandwich	275	grilled chicken salad
149	steel-cut oatmeal	290	honey creme fraiche
205	french toast	328	brunch

We manually labeled food entities in the reviews and removed 32 reviews that do not comprise any food mentioning. The 302 valid reviews were divided into development set I and test set J. A summary of the dataset is provided in Table 2.

Table 2. Summary of review documents and food mentions

	Number of Documents	Number of Food Mentios
Total	302	1042
Set I	200	688
Set J	102	354

EXAMPLE GENERATION AND FEATURE DESIGN

Over the development set I, we generated examples of 1 to 4 word length within each sentence, and pruned examples that are unlikely to be food entity. The pruning rules require 1-word entity to be a noun, and multi-word entity to include only nouns or adjectives. We also constructed a non-food word list, and any potential example that includes any of the word in the list are excluded. The Part-Of-Speech (POS) tag were generated using nltk package.

For each entity example, we generated two types of features: features describing the example (local features) and features examining the context around the phrase (context features). The local features include the phrase itself and attributes such as length of the phrase, its location within the sentence, the type of POS tag it includes, if it has any capital letter, and whether any food ingredient is mentioned. The context features examine POS tag and attributes of the surround words. The final model selected 8 out of 10 local features and 13 out of 52 context features (Table 3).

Table 3. Feature Descriptions

	Feature	Description
Local Features	word_0	The example phrase
	word_0.distBG	Distance from the phrase to beginning of the sentence
	word_0.distEND	Distance from the phrase to end of the sentence
	word_0.distCS	Minimum distance from the phrase to any comma or semicolon within the sentence
	word_0.hasADJ	If the phrase has any adjectives

	word_0.hasCapital	If the phrase has any capital letters
	word_0.hasFood	If the phrase mentions any common food ingredients (we maintain a common food ingredient lists)
	word_0.length	Count of word in the phrase
Context Features (preceding word)	word_l1	Word on the left
	word_l1.hasCapital	If the preceding word has capital letters
	word_l1.isNN	If the preceding word is a Noun
	word_l1.isPRCJ	If the preceding word is a Preposition or Conjuction
	word_l1.postag	Postag of the preceding word
Context Features (following word)	word_r1	Word on the right
	word_r1.isFood	If the following word is a common food ingredient
	word_r1.isNN	If the following word is a Noun
	word_r1.postag	Postag of the following word
Context Features (second-on-the-left and second-on-the-right)	word_l2	Second word from the left
	word_l2.postag	Postag of the second word from the left
	word_r2	Second word from the right
	word_r2.postag	Postag of the second word from the right

CLASSIFICATION

We tested 12 classifiers in sklearn, including decision tree, random forest, support vector machine, linear regression, lasso, logistic regression, adaboost, gaussian process, naive bayes, k-nearest neighbors, mlp classifier, and quadratic discriminant analysis. After doing the first cross validation, we found that random forest (Classifier M) performs the best, which got precision=0.8016, recall=0.5767, and f1=6708. After debugging, the result of final Classifier X (random forest) is listed in Table 4. Since our overall accuracy meets the requirement (90% P and 60% R), we did not perform rule-based post-processing.

Table 4. The precision, recall, and F1 of Classifier X (random forest)

	Precision	**Recall**	**F1**
Set I	0.9012	0.7461	0.8163
Set J	0.9259	0.6849	0.7874

RESULT

Based on our extracted information, we totally created 62 features, and finally selected 21 features to train different models. The best precision appeared in random forest classifier, with precision 92.59%, recall 68.49%, and F1-measure 78.74% on test dataset.