Restaurant Menu Inference
Status Report

Jason Jennings **Nicholas Earley**

Objectives

Develop a system to automatically extract menu items from Yelp restaurant reviews.

Yelp provides businesses the option of linking their menu, and customers of the restaurants often take photographs of the menu that can be viewed as images. This makes it hard to see, at a glance, or search what items a restaurant offers. Additionally, many restaurants have neither a link to their menu nor an uploaded menu image.

Our objective is to infer a restaurant's menu from the text of the restaurant's reviews. This will also allow a user to see the items on the restaurant's menu without visiting an external website, or viewing an image. Additionally, it may be used as a means to provide statistics, such as rating or popularity, for each dish on a restaurant's menu. This can help potential customers make informed decisions about restaurants they choose to attend, and what items they might order.

Deliverables

Website

The primary deliverable will be a website which will allow a user to browse restaurants in the dataset and view the inferred menu items.

Presentation

The presentation will primarily be a demo of the website. Additionally, we may provide examples of data as it goes through the entire process of being classified.

Evaluation

Quantitative

Our classifiers will be evaluated quantitatively, based on common statistics such as error rate on a test set.

We may also compare the results to ground truth results for a small set of restaurants where the ground-truth menu items may be easily available.

Qualitative

Qualitative performance will also be taken into account. Do the menu items inferred from review text seem reasonable?

Modifications of Assigned Tasks

Some changes have been made to the initial plan for developing our system. Based on the recommendations of the professor, we have divided the project into three main components:

Named Entity Recognition

Named Entity Recognition is the task of recognizing which words in a sentence are referring to some entity, in this case menu items or foods.

This task will primarily be the responsibility of Jason Jennings.

Disambiguation

Disambiguation in the context of this project is the task of synthesizing multiple matches that are actually referring to the same named entity.

This task will primarily be the responsibility of Nicholas Earley.

Website

No significant changes have been made to the plan for the website. This will be a joint task by both Jason Jennings and Nicholas Earley.

Implementation Plans

Named Entity Recognition

The plan for the named entity recognition portion of the project involves training a classifier to determine if a sentence is referring to a food item, and if so determining which part of the sentence is the food item. This task still needs significant work, but a basic plan has been developed.

Preprocessing

The data supplied by yelp must be queried to locate the information only relevant to our task.

1. Using the pandas library for python, we can perform a join on the ‘businesses’ file with the ‘reviews’ file on the business_id attribute, with the condition that the business is a restaurant.
2. We will preprocess the data by breaking each review up into individual sentences.

Feature Extraction

The extracted sentences will be fed into the Stanford Parser. The Stanford Parser will allow us to construct syntax trees of the sentences in reviews. By using such a powerful tool, we may select only the features most relevant to our task. Additionally, we will limit the scope of our project to detecting foods in clauses or phrases of sentences that follow a few specific structures that will be hand chosen.

Our features will consist of the following attributes:

Nouns/Adjectives in Noun Phrase – A word vector consisting of the relevant words in a noun phrase.

Verb – A word vector of the specific verb within a verb phrase.

Preposition – A word vector indicating the specific preposition for the prepositional phrase in our sentence.

Prepositional Phrase - A word vector consisting of the nouns in the prepositional phrase in our sentence

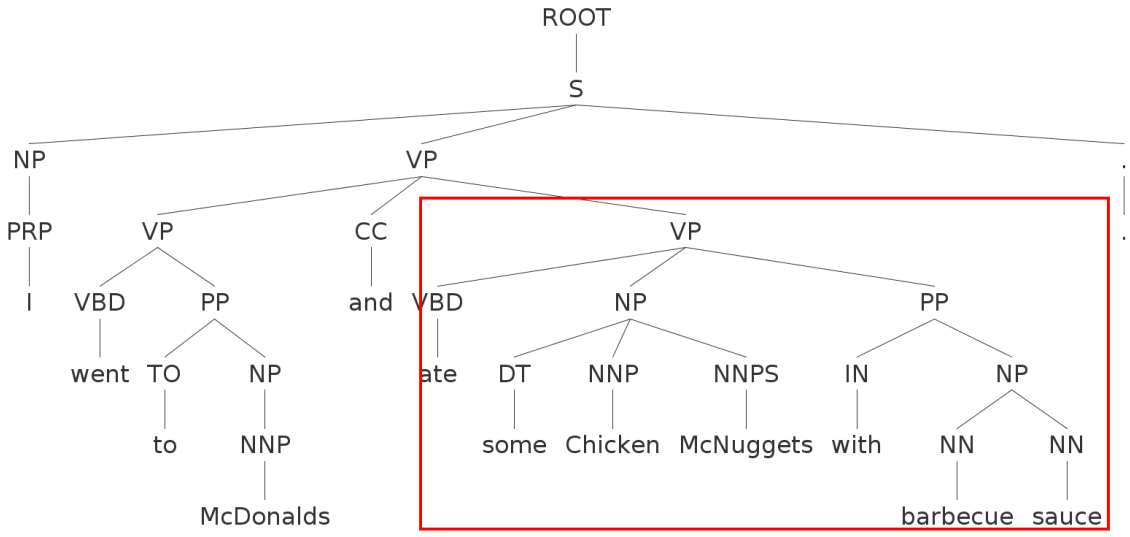
Capitalization – What percentage of the words in the noun phrase had the first letter capitalization.

Intuition: The intuition behind the selected features is that each type of phrase or part of speech can be an indicator of a food. This can be explained by a few examples.

Example 1

Original Sentence: I went to McDonalds and ate some Chicken McNuggets with barbecue sauce.

Syntax diagram:



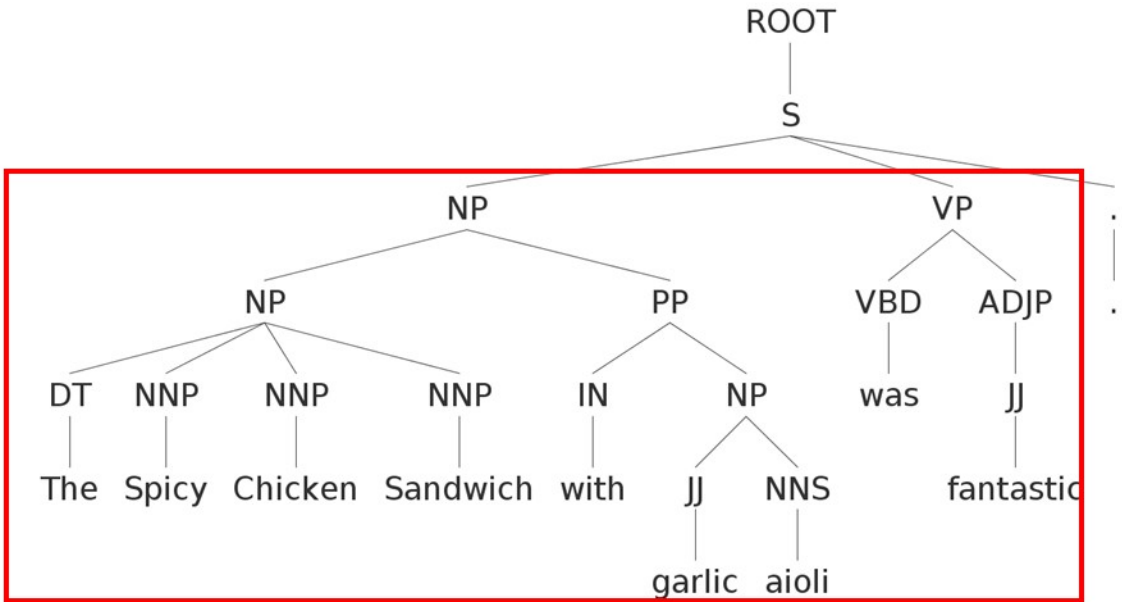
The red box indicates the section of the sentence that would be considered (based on the VBD NP PP structure).

Feature	Value	Analysis
Verb	Ate	'Ate' is a strong indicator of a food

Noun Vector	<Chicken, McNuggets>	Chicken is a good indicator of a food. McNuggets does not have much generalization power, and may not be included based on the nouns selected to form our corpus.
Preposition	With	The preposition 'with' is common, and is likely neutral or at best a marginal indicator of a food.
Object of Preposition - Noun Vector	<barbecue, sauce>	Both of these are very strong indicators of a food.
Capitalization% (of Noun Vector)	100%	This is a good indicator that the item mention is not only a food, but a menu item (which is often a proper noun)

Example 2

Original Sentence: The Spicy Chicken Sandwich with garlic aioli was fantastic.



In this case there is only one clause in the sentence, and all of the information is relevant. Notice this sentence has a different structure than the first example. This time the food is the subject of the sentence.

Feature	Value	Analysis
Verb	Was	Was is a very common verb and is likely neutral at best in terms of classification power
Noun Vector	<Spicy, Chicken, Sandwich>	All three words in this title are strong indicators of food.
Preposition	With	Again, The preposition 'with'

		could strengthen the case for classifying as a food, but alone is not enough information.
Object of Preposition - Noun Vector	<garlic, aioli>	Both of these are very strong indicators of a food. Garlic is likely fairly common, while aioli is somewhat rare.
Capitalization% (of Noun Vector)	100%	Again, the noun vector contains 100% capitals, a strong indicator of a proper noun.

Model Selection

The exact learning algorithm for performing the classification task has not been determined. The current plan is to try Random Forests, Bagged Trees, Naïve Bayes, and Support Vector Machine on the dataset and pick the model which performs the best. However, because these algorithms all have many parameters that may be adjusted, we may limit our scope to just one or two, and instead focus on finding the best set of parameters for those models.

Challenges

Feature Extraction - Extracting the features described above from the syntax tree has proven challenging. We believe this can be overcome with more time and effort.

Data Labeling – Hand labeling enough examples to get good predictive power may be an issue. The task of labeling depends on a working feature extraction implementation.

Model Selection – As mentioned, there are many different machine learning algorithms with different strengths, weaknesses and many parameters to adjust. We may focus on fine-tuning the parameters for one model, rather than trying only one set of parameters on many models.

Current Progress

Disambiguation

Disambiguation in the context of this project is the task of synthesizing multiple matches that are actually referring to the same named entity.

Approach

We have a few ideas for the task of disambiguation. The main idea behind our approach is to use similarity measures such as cosine similarity as well as edit distance to identify phrases that are similar to a given query phrase. If the similarity is higher than some threshold (the threshold may be learned or hand-selected), we will say these two items refer to the same named entity.

Other Thoughts

Some other approaches may involve using libraries that perform spell checking and correction to aid in the task of disambiguation, since misspelled words could be a source of many duplicate foods in our inferred menu. Stopword removal and stemming will likely improve the performance as well.

Challenges
Implementation has been difficult because the task depends on the named entity recognition task, which has not been completed.

Current Progress
Initial idea has been researched. Baseline implementation is in progress based on cosine similarity, but needs results from named entity recognition to be evaluated in a meaningful way.