Word2Mouth: A Top Dish Recommender System

Abhirup Mukherjee
amukherjee@umass.edu

Mithra Muthukrishnan
mmuthukrishnan@umass.edu

Aditya Narula
anarula@umass.edu

Aishwarya Turuvekere
aturuvekere@umass.edu

1 Introduction

Yelp provides a convenient way to discover restaurants with top reviews by users. However, besides these reviews, most customers are also interested in information about the businesses themselves, such as menus for restaurants. Customers also often require information about a restaurant's most popular dishes in order to make a decision on what to order, especially if it's their first time at the establishment. We noticed that most restaurants on Yelp don't have menu items or recommended dishes. Most of this information is buried in large amounts of user review data. This entails more work for users, including, but not limited to, reading and parsing a sufficient amount of reviews by other users to get an idea of what the popular dishes are at an establishment and what these dishes look like.

We propose a system to automatically detect and extract dishes from reviews and create a list of popular dishes for each restaurant. Our solution aims to be independent of the availability of menus, which are often not available or difficult to source. The only requirement would be a large enough corpus of reviews to allow our system to accurately extract the top dishes information. The problem can be broken down into named entity recognition, where we treat food as a category of entity. This system would be relevant to all types of restaurant goers, from the casual diners who are new to the city or are exploring the area for the first time, to avid food connoisseurs trying out the food at a well rated restaurant for the first time. This would also be valuable to restaurant owners to keep track of the most sold dishes, without having to sift through thousands of reviews.

2 Related work

Yelp has recently implemented a 'Popular dishes' section on their website to take the guesswork out of ordering at different restaurants. On the Yelp app, you can now see the most popular dishes along with dish names, photos and review snippets. Their blog (Geddie, 2018) summarizes the methodology as an algorithm which looks at the number of times a dish has been mentioned in reviews, the star rating for the business among a number of other things. However, Yelp does not provide this list for restaurants without menus, even if the restaurant has a large number of reviews, and is reasonably popular. There is also no direct access to a detailed overview of the methodology and process followed by the Yelp engineering team; no papers are available related to the same. During our research we also found that currently there are no supporting applications either which aggregate popular dishes for the restaurants listed on Yelp. The user would now be required to go though several different websites, consult others and spend a lot of effort reading a sufficient number of reviews amongs other things to get a decent idea about what is trending, which is a tedious task.

We found a system (Vechtomova, 2014) which performs automatic extraction of multi-word units from user reviews. The system employs a semi-supervised approach, using a list of seed phrases that indicate the category of interest. The system proposes a method which first generates candidate multi-word aspects by performing POS tagging on the review, and then merging adjacent words by looking at the normalized pointwise mutual information. The system compares many distributional similarity measures and uses a document retrieval function, BM25.

The main task in our project, identification of

dishes in a semi structured body of text, essentially boils down to a named entity recognition problem. We will now look at some previous work on named entity recognition using different models.

One way to represent the NER model is through a neural architecture. (Lample et al., 2016) discusses two different types of architectures. Both models do not use any language specific resources, but rely on two sources of information about the word: character-based word representations learned from a supervised corpus and unsupervised word representations learned from an unannotated corpus. To implement an NER, they compare the performance of two models: a bi-directional LSTM with a CRF above it and a new model based on transition-based parsing. Their experiments found that the LSTM-CRF hybrid model was superior by a good margin. We think the bi-directional LSTM-CRF hybrid model would be a good match for our task. We plan to implement this model and incorporate modifications to fit the task at hand.

3 Approach

Our approach can be split into four main steps: extracting dishes/food names from the reviews, aggregating the dishes to form a menu, associating a sentiment or opinion for each dish in the menu (based on the review), and finally creating a popularity list based on the sentiments recorded for each menu item. To identify food dishes we are using an NER system. We will compare two different implementations: spaCy model trained on our annotated data and an implementation bi-directional LSTM-CRF hybrid to see which performs better for our use case. We plan to use different similarity metrics to create a distinct set of menu items. The sentiment analysis tool in Stanford CoreNLP will be used to determine the sentiment to each review. We plan to rank the food items higher if we see a larger number of positive reviews associated with it. The calculated ratings would be used to get the popularity list.

Our baseline - A simple way to tag all food items would be to have a large dictionary with popular food names taken from (Bender, 2014) and to simply look for these items in the reviews. We would first be tokenizing and normalizing the reviews before looking for an absolute match for the dictionary items. This naive approach might work well with popular items, but we foresee a few drawbacks. The dictionary might not be comprehensive enough to enumerate foods from all cuisines. It would also be hard to include dishes with non-english names, like bún bò huẽ or baklava, with the right spellings. Restaurants tend to include adjectives like crispy, cheesy or special and this baseline would not be able to capture those.

3.1 Milestones & Schedule

1. Acquire and preprocess data (2 weeks)
2. Build models for task (3 weeks - done in parallel with task 1)
3. Write progress report (1 week)
4. Output and Error Analysis for Model (2 weeks)
5. Work on final report and presentation (2 weeks)

4 Data

We will use the Yelp Open Dataset for the reviews, which approximately contains 2.7 million reviews for 86K businesses written by 687K different users. The dataset includes five JSON formatted objects containing businesses, users, and review data. The business object holds information such as business name, description, location, category, rating etc. The review object contains star rating, review text, user information and usefulness voting among other details. Yelp automatically filters out spam reviews, so we will only use reviews from Yelp for this project. However, the data set does not provide menus for restaurants. To account for this missing data, we will use a scraper to get the menu items for validation purposes. For the out-of-the-box NER model and the bi-directional LSTM-CRF hybrid, we would just require a amount of the training data, which we will be hand annotating. This task will be broken up among all members of the group. For the baseline, we will use a food dictionary,(Bender, 2014) Oxford: An A-Z of Food and Drink.

5 Tools

We will be using the following tools/libraries for our project -

1. Helena/ROUSILLON (Chasins et al., 2018) - a scraper and an automator to create a validation set

2. Stanford CoreNLP (Manning et al., 2014) - for tokenization, normalization and sentiment analysis

3. spaCy - an open-source library to build a NER model, trained on our annotations

4. TensorFlow (Abadi et al., 2016) - a machine learning library, will be used to implement the NER

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., and Zheng, X. (2016). Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283.

Bender, D. A. (2014). A dictionary of food and nutrition. In *A Dictionary of Food and Nutrition*.

Chasins, S. E., Mueller, M., and Bodik, R. (2018). Rousillon: Scraping distributed hierarchical web data. In *The 31st Annual ACM Symposium on User Interface Software and Technology*, pages 963–975. ACM.

Geddie, M. (2018). Introducing popular dishes on yelp: Taking the guesswork out of what to order. In *Introducing Popular Dishes on Yelp: Taking the Guesswork Out of What to Order*. Yelp.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270. Association for Computational Linguistics.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Inc, P., Bethard, S. J., and Mcclosky, D. (2014). The stanford corenlp natural language processing toolkit. In *In ACL, System Demonstrations*.

Vechtomova, O. (2014). A method for automatic extraction of multiword units representing business aspects from user reviews. *J. Assoc. Inf. Sci. Technol.*, 65(7):1463–1477.