Kasam Dhakal

CS 598 Data Mining Capstone - Task 3

Dish Recognition

Date: 6/12/2024

[kasamdh/CS598Capstone (github.com)](https://github.com/kasamdh/CS598Capstone)

kdhakal2@illinois.edu

**Dish Recognition**

The objective of this assignment is to analyze the dataset to uncover the prevalent or well-known dishes of a specific cuisine. Often, when experiencing a new cuisine, the variety of dishes it offers is unfamiliar. Our aim is to determine the available dishes of a cuisine by developing a dish identification system.

**Task 3.1: Manual Tagging**

In this task, I opted to concentrate on Mexican cuisine because it had the fewest candidate dish names, and I prefer to minimize manual effort. Following the instructions, I quickly reviewed the 130 dish names and created a column with my manual labels, which I considered the "truth." I generally labeled any phrase that was not a dish name as 0, even if the phrase could be associated with "Mexican" cuisine, as the overview emphasizes identifying specific dishes available in a cuisine. However, if a dish could be considered part of Mexican cuisine (e.g., Tacos al Pastor or Chiles Rellenos), I labeled it as 1. Following the suggestions, I removed all identified false positives and corrected the labels for all false negatives. The result of this task was a file named Mexican\_updated.label, which I placed in the /data directory for the SegPhrase section.

Remove False Positive of Mexican Cuisine

1. food poisoning 1
2. salad bar 1
3. fast food 1
4. fast food chain 1
5. fast food restaurants 1
6. coca cola 1
7. taco bell 1
8. del taco 1
9. el pollo loco 1
10. panda express 1
11. die hard 1
12. kick ass 1
13. sammy hagar 1
14. hip hop 1
15. classic rock 1
16. service stars 1
17. wi fi 1
18. iron chef 1
19. language barrier 1
20. saving grace 1
21. date night 1
22. spring break 1
23. super bowl 1
24. green beans 1
25. sticky rice 1
26. sour cream 1
27. guilty pleasure 1

**Task 3.2 Mining Additional Names**

I opted for SegPhrase over ToPMine or word2vec primarily because it is recommended as the state-of-the-art method in the instructions. Moreover, it allows me to effectively use my manual efforts from Task 3.1. According to the instructions, "SegPhrase has a classifier to assign a quality score to each phrase candidate based on their statistical features. The classification procedure will be enhanced by phrasal segmentation results. These two parts could mutually enhance each other." While ToPMine is suitable for phrase mining without training data, incorporating a small set of training data can improve the quality of phrase mining, which is why using SegPhrase with our labeled phrases and a constructed Knowledgebase is beneficial.

SegPhrase is specifically designed for phrase mining with small training datasets, making it perfect for this problem. It utilizes labels (such as our human-annotated labels or those from a knowledgebase) to assess the quality of a phrase. SegPhrase then uses the random forest algorithm for classification, building models that differentiate between high-quality and low-quality phrases. Additionally, it employs phrasal segmentation to identify the most appropriate phrases by maximizing likelihood. Overall, SegPhrase is well-suited for this task because it can handle a small training set, perform feature extraction, and determine the informativeness or quality of a phrase. It effectively uses IDF and mutual information to identify relevant dishes and filter out irrelevant ones.

**Setup**

To set up SegPhrase, I first cloned the repository from <https://github.com/shangjingbo1226/SegPhrase>. Used Homebrew to install gcc49, and modified the g++ variable in my Makefil to point to my gcc path (export CXX = /usr/local/Cellar/gcc@4.9/4.9.4\_1/bin/g++-4.9). I also created a conda environment for Python 2.7 and installed scikit-learn and nltk, which I activate when running the shell scripts. I then generated a knowledgebase of common Mexican dishes from <https://en.wikipedia.org/wiki/List_of_Mexican_dishes>, added 250 dish names from the article, and appended them to /data/EN/wiki\_quality.txt from the AutoPhrase repo to create /data/wiki\_labels\_quality\_append\_MexicanDish.txt. Next, I modified the provided train\_toy.sh script with parameters such as RAW\_TEXT='data/Mexican.txt' from Task 2, containing all Yelp reviews related to Mexican cuisine, and KNOWLEDGE\_BASE='data/wiki\_labels\_quality\_append\_MexicanDish.txt'. Finally, I ran the modified script with ./train\_MexicanDish.sh.

**First**

Initially, I used the data labels mentioned above. Although SegPhrase performed reasonably well in generating output phrases, it still produced several false positives and had difficulty with ambiguous categories where the phrase pertained to Mexican culture rather than specifically to cuisine.

|  |  |
| --- | --- |
| fried egg | 0.998857168 |
| in n out | 0.99792519 |
| triple sec | 0.997713903 |
| mexican food | 0.997379549 |
| service stars | 0.997351527 |
| food poisoning | 0.997261925 |
| self service | 0.99715519 |
| salad bar | 0.996804812 |
| beef tongue | 0.996804812 |
| main course | 0.996391304 |
| desert inn | 0.996236284 |
| wi fi | 0.995780933 |
| mexican american | 0.995741925 |
| arnold palmer | 0.995741925 |
| los olivos | 0.995527236 |
| birthday girl | 0.995261925 |
| diners drive ins and dives | 0.995075258 |
| red snapper | 0.994941925 |
| dive bar | 0.994107588 |
| van buren | 0.993636662 |
| del taco | 0.991797101 |
| tex mex | 0.99135218 |
| refried beans | 0.990149853 |
| carne asada | 0.98840519 |
| die hard | 0.986716284 |
| celebrity chef | 0.985939327 |
| green beans | 0.985713903 |
| coca cola | 0.985276062 |
| san diego | 0.985276062 |
| kick ass | 0.97840519 |
| tater tots | 0.975728039 |
| skirt steak | 0.967741925 |
| tap water | 0.967261925 |
| food chain | 0.965939327 |
| filet mignon | 0.965276062 |
| iron chef | 0.963941898 |
| sour cream | 0.95840519 |
| fast food | 0.95840519 |
| stir fry | 0.95840519 |
| sticky rice | 0.956391304 |

False Positive: 10

**Second 2**

|  |  |
| --- | --- |
| refried beans | 0.999156153 |
| tex mex | 0.999156153 |
| mole sauce | 0.999156153 |
| mexican cuisine | 0.997890711 |
| carne asada | 0.99674297 |
| food allergies | 0.996644708 |
| fast casual | 0.995076304 |
| taco bell | 0.993038088 |
| carnitas tacos | 0.993038088 |
| comfort food | 0.989795259 |
| la bamba | 0.986854083 |
| language barrier | 0.984043204 |
| el hefe | 0.982393454 |
| maria maria | 0.979565057 |
| el segundo | 0.979370539 |
| frida kahlo | 0.97706684 |
| classic rock | 0.97684241 |
| sea bass | 0.976455586 |
| blind pig | 0.976359002 |
| apache junction | 0.973038088 |
| taco bell | 0.969384489 |
| tater tots | 0.967324741 |
| tortilla soup | 0.966972855 |
| guacamole | 0.966817936 |
| heart attack | 0.962506452 |
| iron chef | 0.96123202 |
| green onions | 0.95382734 |
| diet coke | 0.939596421 |
| newport beach | 0.895114952 |
| fried chicken | 0.894532005 |
| potato chips | 0.894458144 |
| flank steak | 0.894458144 |
| bottled water | 0.876854083 |
| fast food chain | 0.868971602 |
| roast beef | 0.866854083 |
| guy fieri | 0.850478774 |
| cream cheese | 0.845609219 |
| fast food restaurant | 0.84359657 |
| noodle soup | 0.843173786 |
| rice pudding | 0.838366902 |

False Negative: 8

To enhance my results, I generated a new file named Mexican\_updated2.label and implemented two adjustments: rather than simply removing them, I corrected the false positive dish names after noticing they were recurring. Additionally, I augmented the dataset by including new positive labels derived from the top 20 results of a Google search for "Mexican dishes." These changes aimed to refine the accuracy of SegPhrase in identifying relevant Mexican cuisine terms while reducing ambiguous or unrelated phrases.

Tacos, Enchiladas, Guacamole, Chiles Rellenos, Tamales, Quesadillas, Pozole, Mole Poblano, Sopes, Tostadas, Fajitas, Burritos, Chilaquiles, Carne Asada, Sopa de Tortilla (Tortilla Soup), Elote (Mexican Street Corn), Huevos Rancheros, Barbacoa, Carnitas, Ceviche.

I changed DATA\_LABEL='data/Mexican\_updated2.label' and ran ./train\_MexicanDish.sh again. Upon reviewing the top 50 results, there wasn't a noticeable improvement, but we did discover a few additional dishes that were missed initially. I think I could continue refining my results by manually correcting the output to address false positives. This iterative approach, supported by SegPhrase, allows for continuous improvement through feedback, which is one of its strengths.

**References:**

1. El-Kishky, A., et al. (2014). Scalable topical phrase mining from text corpora. Proceedings of the VLDB Endowment, 8(3): 305-316.
2. Liu\*, J., Shang\*, J., Wang, C., Ren, X., & Han, J. (2015, May). [Mining quality phrases from massive text corpora](http://jialu.cs.illinois.edu/paper/sigmod2015-liu.pdf), Proceedings of 2015 ACM SIGMOD International Conference on Management of Data (SIGMOD'15). Melbourne, Australia. (\* equally contributed)
3. Goldberg, Y., & Levy, O. (2014). word2vec explained: Deriving Mikolov et al.'s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722.
4. <https://en.wikipedia.org/wiki/List_of_Mexican_dishes>
5. <https://github.com/shangjingbo1226/SegPhrase>