Task 3: Dish Recognition

Xiaoming Ji (xj9)

Overview

The goal of this task is to mine the data set to discover the common/popular dishes of a particular cuisine. Typically, when you go to try a new cuisine, you don't know beforehand the types of dishes that are available for that cuisine. For this task, we would like to identify the dishes that are available for a cuisine by building a dish recognizer.

For comparison, two approaches: SegPhrase and AutoPhrase are used to extract dishes from **Chinese** cuisine. SegPhrase and AutoPhrase are algorithms developed by UIUC and achieved SOTA performance on mining quality phrases. SegPhrase needs a small amount of manually tagged samples. While AutoPhrase reduced human effort by leveraging existing general knowledge bases.

We can see both approaches achieved good results but AutoPhrase performs better in term of finding new phrases with less effort.

Data Preparation

Cuisine corpus

Extract all (38715) Chinese restaurants related reviews from the whole dataset using script *processYelpRestaurants.py*.

Tagging

To support SegPhrase, we produced 2 label files by modifying the provided *Chinese.label* file,

1. *Chinese.updated.label*:
    a. Removed 128 false positive non-dish name phrases. E.g., los angeles, soft serve.
    b. Changed 6 false negative dish name phrases to a positive label. E.g., general tso's chicken, wonton strips.
2. *Chinese.update.plus.label*:
    a. Based on *Chinese.updated.label*, we added 26 new positive dish name phrases. These phrases are mined from AutoPhrase. Note: adding negative samples don't seem to improve the results.

Dish Name Knowledge Base

AutoPhrase leverages the existing high-quality phrases, as available from Wikipedia or other trusted sources. Since we are mining dish names, we find 132 Chinese dish names from List of Chinese Dishes and produced 2 knowledge base files,

1. *Chinese_dish_quality.txt*: only contains these wiki Chinese dish names.
2. *wiki_chinese_dish_quality.txt*: we append these wiki Chinese dish names to wiki_quality.txt (found in <AutoPhrase installation>/data/EN)

Dish Name Mining

SegPhrase

SegPhrase was run with following parameters,

- SUPPORT_THRESHOLD=10
- DISCARD_RATIO=0.00
- MAX_ITERATION=5
- ALPHA=0.85
- WORDNET_NOUN=0

See https://github.com/shangjingbo1226/SegPhrase for explanations on these parameters.

3 label files are used to mine new Chinese dish names.

1. Chinese.label: the provided/original label file. Shown as **Seg.Chinese**
2. *Chinese.updated.label* : Shown as **Seg.ChineseUpdated**
3. *Chinese.update.plus.label*: Shown as **Seg.ChineseUpdated+**

AutoPhrase

AutoPhrase was run with the following parameters,

- ENABLE_POS_TAGGING=1, AutoPhrase will utilize the POS tagging in the phrase mining.
- MIN_SUP=10, a hard threshold of raw frequency for frequent phrase mining.

3 dish name knowledge base files are used to mine new Chinese dish names.

1. *wiki_quality.txt*: the wiki quality phases found in AutoPhrase installation. Shown as **Auto.Wiki**.
2. *Chinese_dish_quality.txt*: Shown as **Auto.ChineseDish**.
3. *wiki_chinese_dish_quality.txt*: Shown as **Auto.WikiChineseDish**.

Results

The top 40 new Chinese dish names mined by each approach are listed in following figure. New means we don't count the dish names in the label/knowledge base files. Wrong dish names are marked in red. Some phrases are marked in blue and they are food but not Chinese/Asian dish names.

Seg.Chinese	Seg.ChineseUpdated	Seg.ChineseUpdated+	Auto.Wiki	Auto.ChineseDish	Auto.WikiChineseDish
chow fun	soy sauce	orange chicken	bone marrow	chow mein	tom yum goong
decent chinese food	orange chicken	panda express	bento box	ice cream	bento box
chinese places	fountain hills	banh mi	bean sprout	dim sum	sesame seed
long beans	pepper squid	lunch special	singapore sling	bean curd	bean sprout
thai restaurant	egg rolls	pan fried	napa cabbage	chow fun	olive garden
orange tofu	beef chow fun	papaya salad	refried beans	fried rice	pork belly
chicken curry	chow fun	teriyaki chicken	sugar cane	peking duck	roti canai
high prices	mongolian beef	roast pork	orchid garden	kung pao chicken	spare ribs
coconut shrimp	happy hour	egg flower soup	turnip cake	pad thai	fortune cookie
lettuce wraps	crab puffs	fountain hills	roti canai	beef chow fun	mashed potatoes
egg noodles	lettuce wraps	general tsao chicken	credit cards	siu mai	mung bean
lunch break	black bean sauce	thai coconut curry	fountain hills	hot pot	ahi tuna
teriyaki bowl	monte carlo	bell pepper	pupu platter	beef noodle soup	napa cabbage
cathay house	jade red	soy sauce	sin city	shaved ice	orchid garden
bento box	lotus leaf	beef noodle soup	olive garden	pork belly	french fries
red peppers	panda express	clay pot	west lake	lo mein	bone marrow
ice tea	bbq pork	monte carlo	tom kha	noodle soup	credit cards
chili paste	egg drop soup	green onion	scrambled eggs	bbq pork	eureka casino
iron chef	wonton noodle soup	lettuce wraps	peanut butter	sea bass	fountain hills
wor wonton soup	teriyaki chicken	chile relleno	mashed potatoes	dan dan noodles	pupu platter
china tango	beef noodle soup	crab puffs	monterey park	won ton	caesar's palace
green bean	hot sour soup	chili paste	iceberg lettuce	green tea	shark fin
roast pork	chinese restaurants	pork belly blt	cheesecake factory	pf changs	boulder city
black sesame	sesame chicken	lunch specials	golden gate	honey walnut shrimp	singapore sling
chinese joint	crab rangoon	egg rolls	bok choi	orange chicken	peanut butter
online menu	pan fried	black beans	filet mignon	green beans	sin city
lotus leaf	thai iced	kung fu	lotus root	panda express	string bean
crab ragoons	lunch specials	rice cakes	almond pudding	lo mien	cheesecake factory
big buddha	oyster sauce	chili sauce	tuna tataki	mapo tofu	asian cuisine
bbq ribs	jade red chicken	hot sour soup	mung bean	char siu	tuna tataki
singapore noodles	chicken curry	asian cuisine	chiu chow	egg drop soup	lemon grass
green peppers	singapore noodles	fried calamari	ahi tuna	lettuce wraps	san jose
boba tea	egg foo	red bean	tom yum goong	egg drop	west lake
massaman curry	china king	black sesame	shopping center	pork chop	snow pea
fish fillet	won ton soup	lunch break	shui mai	roast duck	chocolate mousse
tea pot	lunch special	string bean	cheddar cheese	xiao long bao	filet mignon
udon noodles	jerk fried rice	honey seared chicken	french fries	iced tea	sugar cane
orchids garden	clay pot	dan dan noodle	pollo diablo	sesame chicken	iced tea
seating area	black pepper	crab legs	sesame seeds	shark fin soup	caesars palace

Conclusion

By checking the results, we can see both SegPhrase and AutoPhrase can find good quality of Chinese dish names. To conclude,

- Dish names are most commonly mentioned phrases in restaurant reviews. Even without specific label/knowledge base and just use phrases from Wiki, we see large portion of phrases mined by **Seg.Chinese** (Chinese.label are generated from wiki phrases) and **Auto.Wiki** are quality Chinese dish names.

- AutoPhrase is very efficient on finding new dish names. No manual tagging is needed. Although the knowledge base of 132 Chinese dish names is small, the results are still very impressive.
- **Auto.ChineseDish** (only use Chinese dish names as the quality phrases) is the most effective approach to find new Chinese dish names. Only 2 errors found compared with 10+ errors as of produced by other approaches. For further improvement, we can increase the size of the knowledge base.
- **Seg.ChineseUpdated+** performs a slightly better than **Seg.ChineseUpdated** in terms of error rates (10 errors vs. 12 errors). It seems adding more positive tagged samples won't improve performance significantly. In addition, our testing shows adding negative tagged samples won't improve performance at all.