Task 3 - Dish Recognition

Task 3.1 Manual Tagging

Objective

Refine the label list of one cuisine in given file

Tools

SegPhrase

For this task, there is no labels that can be used as reference to train the input file, therefore I applied the default label info in SegPhrase, after that I could get the labels of Chinese cuisine reviews and tips.
Based on these information, I could make further inference about missing dishes.

Below is what I did for this task:

1. Since dish names are mentioned in both review file and tips file, I classified reviews and tips of all restaurants by category, for instance all Chinese restaurants' reviews and tips, then record them in a file named Chinese-reviews.txt

2. Train this file using SegPhrase with default setting (default label), get labels marked with 0 and 1 in wiki.label.auto file

3. As mentioned in the instruction, some of the provided labels could be wrong, so I need to merge labels in wiki.label.auto with provided labels into task1_label file in order to find the difference between what I found and what I was given.

   A. Loop all the labels in provided files to find potential conflict
      1. If the label not exist in wiki.label.auto file, I will skip it as this is a safer way comparing with switching value from 0 to 1 or vice versa.
      2. If the label exists in wiki.label.auto file, I'll include it in output file. The corresponding value will adopt the one from wiki.label.auto result.

B. Check missing labels by looping all labels in wiki.label.auto result, since it's not many I did some manual filtering work by including only those related to dish phrases.

Then submit the final task1_label file to auto grader.

This file can be found here:
https://drive.google.com/file/d/0B1aRlNAJ0UZELUd1blo4ZklNRE0/view?usp=sharing

Other possibilities

Actually I tried to use the wiki.label.auto file from SegPhrase directly, however the result is not optimal as it contains some unrelated phrases.

Analysis

The result in wiki.label.auto includes some unusual phrases such as "los angeles", "south california", chances are these phrases are mentioned by people when they are giving reviews in that it would be more useful if specific address of restaurant can be listed. However they should be excluded from dish name labels.

Also some words such as "bra" which is totally irrelevant to dish names are listed, I just deleted these items to make sure the final file is purely cohesive.

Task 3.2 Mining Additional Dish Names

Objective

Expand missing dish names

Tools

ToPMine, SegPhrase

Here is what I did:

1. Train the reviews and tips of Chinese category using SegPhrase
   Below is the setting for train.sh
   RAW_TEXT = 'Chinese-reviews.txt'

AUTO_LABEL = 0 //to force it to use the label file obtained got from task 1
DATA_LABEL = task1_label_file
KNOWLEDGE_BASE, KNOWLEDGE_BASE_LARGE not changed
SUPPORT_THRESHOLD = 5 //adjust to 5 to make sure some rare dish names will not be filtered
DISCARD_RATIO = 0.05
MAX_ITERATION = 5
NEED_UNIGRAM = 0
ALPHA = 0.85

The output salient.csv file contains phrases according to their possibilities in descending order, I replaced all "_" in label with space, then picked top 5000 lines as first part

2. Train the Chinese-reviews.txt file using ToPMine
   Below is the setting for run.sh
   minsup = 5
   maxPattern = 5
   topicModel = 2
   numTopics = 10
   gibbsSamplingIterations = 1000
   thresh = 5
   optimizationBurnIn = 100
   alpha = 2
   optimizationInterval = 50

From the topPhrases.txt in output folder, I picked the top 3000 lines as second part

3. Combine labels of first and second part to the task2_output file.

4. Finally, I removed duplicated dishes in task2_output file to get the final result which contains around 6,900 items.

The topPhrase file: https://drive.google.com/file/d/0B1aRlNAJ0UZETHgwblh4ZG1HWTQ/view?usp=sharing

Salient file: https://drive.google.com/file/d/0B1aRlNAJ0UZERXBXcS0zTDJ5REk/view?usp=sharing

task2_output file: https://drive.google.com/file/d/0B1aRlNAJ0UZEVnJXQkhfX193V0E/view?usp=sharing

Other possibilities

At the beginning, I tried with SegPhrase , TopMine and task1JavaTool separately, however each time I could only get 5.0

Then I found that there is hint in the feedback after submission to use both SegPhrase and ToPMine results. That's how I though of the way to combine result from SegPhrase and ToPMine and finally get full marks.

Analysis

Definitely there are differences between the output of these tools, but more importantly, they can compensate each other to find out more hidden dish names. Also for such kind of mining task, minimum threshold needs to be adjust if you want to find out more rare cases.

Still there are many "strange" items among the final result like "alma school", "look forward", this time I just left as it is since there are too many such phrases to be filtered.