

Foundations of Natural Language Processing

Peking University, 2024

Lab 4 Project 2: Dish Name Recognition

1. Directions

Please first read the **general instructions** of Lab 4.

If you choose this project,

- Please submit your homework as a zip file through **Course**, which should include one report in PDF and your source code in Python.
- The report should be generated from the **LaTeX template** provided in Lab 1. Please include the score you achieved on the leaderboard in the report.
- The code should be paired with a README file describing dependencies, code structures, etc.
- There is no need to submit the data you used and the model weights. Your grade will be based on the contents of the report and the source code.

We will not simply grade your homework based on the model performance, but consider **the models you use**, the **novelty** of your method, the **workload**, and the **analysis** in your report.

If you graduate this summer, given that you have less time to complete the project, we will apply a more relaxed grading scale.

2. Task Description

In the dish name recognition task, given the instructions of a recipe, models are required to predict the dish name, which is often composed of up to three kinds of components: food, actions, and flavor.

For example, the dish name '红烧肉' consists of the flavor '红烧' and the food '肉', and the dish name '西红柿炒鸡蛋' consists of the food '西红柿' and '鸡蛋' and the action '炒'.

The recipes are collected from xiachufang.com. The training set contains 187,520 recipes with 3,109 distinct dishes. The test set contains 1,000 recipes. Note that almost none of the dish names in the test set are present in the training set, but all the components of the dish names appear in the training set. For example, if there is a recipe for the dish ‘红烧牛腩’ in the test set, the training set is unlikely to contain other recipes for the dish ‘红烧牛腩’, but is guaranteed to contain recipes for other dishes related to ‘红烧’ and ‘牛腩’ (like ‘红烧肉’ and ‘卤牛腩’). This requires models to learn knowledge about the components, and compose them into unseen dish names.

There is **no constraint** on the method you use. You can implement your own model from scratch, finetune on (large) language models, or use APIs. If you would like to use APIs, please use the Deepseek API as mentioned in the general instructions. Please clearly describe the method (including the prompt for API) you use in the report.

For details of data and evaluation, please check the Kaggle competition (<https://www.kaggle.com/t/601026e0eaff4aab899638a490223a3d>).

3. Resources

The task inherits from the paper *DiNeR: A Large Realistic Dataset for Evaluating Compositional Generalization* (<https://aclanthology.org/2023.emnlp-main.924/>). You may read the paper for a high-level idea of **how to address the task in the aspect of compositional generalization**, but **do not** copy the methods from the paper.