Title
Project Proposal for NLP Course, Winter 2022

Maciej Chrabaszcz
Warsaw University of Technology
maciej.chrabaszcz.stud
Aleksander Kozłowski
Warsaw University of Technology
aleksander.kozlowski.stud@pw.edu.pl

supervisor: Anna Wróblewska
Warsaw University of Technology
anna.wroblewska1@pw.edu.pl

Abstract

The task aims to enrich the RecipeNLG dataset. We want to include additional data like dietary tags and nutritional values. Dataset as is misses few important data which would enrich it and allow for creation of more models related to recipes data. We want to use other datasets with tags to enrich RecipeNLG with dietary tags ex. ”Vegan” which could allow for conditional recipe generation. Another aspect is getting quantities, state and name of products used in recipes, for which we want to use NER models. Finally we want to include more nutritional facts about recipes by using products and their quantities found in ingriedients.

1 Scientific goal

The RecipeNLG dataset includes text describing ingredients, instruction and some products without their quantity. One can see that it would be beneficial to add additional data. Our challenge is to add dietary tags so we would know which recipes are vegan diary-free etc. Knowing which products are used in recipe could sometimes not be enough for specific problems. Because of that will include at least quantity of products used in recipe. Another interesting information that one could want to find is nutritional information about recipe, which we will try to find using online databases related to nutritional values of ingredients used in recipes.

2 Significance of the project

RecipeNLG is one of the biggest datasets related to recipes. After adding this additional data this dataset would become useful not only for generation but some other problems related to recipes text. With this additional data it would be possible to create conditional generative models for recipes given their dietary tag or products quantities. Having dietary tags it is possible to train classification models for dietary tags given ingredients.

3 Concept and work plan

We are planing to use collected data realated to recipes ingredients to create automatic dietary tagger using techniques like TF-IDF, BERT etc. We are going to test multiple text embeddings to see how they work for classification. We will also compare language models trained on recipes data and data scraped from web. We will also look at NER models pre-trained on recipes data to extract products and their quantities.

4 Approach research methodology

On this project we will be using python with libraries related to NLP and Machine Learning ex. pytorch, transformers, nltk, spacy, scikt-learn. For our computations we are planning to use free resoursces from google (Google Colabolatory and Google Drive). We will be testing our methods on preselected part of RecipeNLG which models will never see during training. For classification we will report most important measures related to this task ex. ROC AUC, F1, accuracy, precision, recall. For NER we will look at what percentage of foods and quantities were retrieved from recipes and how many of them are wrong.

References

Bien, M., Gilski, M., Maciejewska, M., Taisner, W., Wisniewski, D., & Lawrynowicz, A. (2020). RecipeNLG: A Cooking Recipes Dataset for Semi-Structured Text Generation.

Ania, W., Agnieszka, K., Maciej, P., Dawid, W., Witold, S., Agnieszka, Ł. (2022). TASTEset – Recipe Dataset and Food Entities Recognition Benchmark. https://doi.org/10.48550/arxiv.2204.07775

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. https://doi.org/10.18653/v1/2020.acl-main.747

Britto, L., Pacífico, L., Oliveira, E., & Ludermir, T. (2020). A Cooking Recipe Multi-Label Classification Approach for Food Restriction Identification. https://doi.org/10.5753/eniac.2020.12133

Nirav, D., Devansh, B., Ganesh, B. (2020). A Named Entity Based Approach to Model Recipes. https://doi.org/10.48550/arxiv.2004.12184

Howard, Jeremy, and Sebastian Ruder. ”Universal language model fine-tuning for text classification.” arXiv preprint arXiv:1801.06146 (2018).

Li, Shuyang Li, Yufei Ni, Jianmo McAuley, Julian. (2021). SHARE: a System for Hierarchical Assistive Recipe Editing.

FoodOn: A farm to fork ontology https://foodon.org/

https://www.kaggle.com/datasets/shuyangli94/foodcom-recipes-with-search-terms-and-tags