Project Ideas Using a Structured Recipe Dataset

Your Name

Dataset Overview

We work with a structured recipe dataset obtained via the Foodoscope RecipeDB API. Each recipe contains:

- A list of ingredients
- Step-by-step cooking instructions
- Structured constraints such as cooking time, calorie information, and dietary labels

This structure enables multiple precise Natural Language Processing (NLP) tasks that go beyond generic text generation.

Idea 1: Constraint-Aware Recipe Instruction Generation

Problem. Standard recipe generation models often ignore hard constraints such as cooking time limits, calorie ranges, or dietary requirements.

Research Question. Can language models generate coherent cooking instructions while strictly satisfying numeric and categorical constraints?

Task Definition.

- **Input:** Ingredient list + constraints (e.g., cook_time ≤ 30, calories ≤ 400, diet = vegetarian)
- **Output:** Step-by-step cooking instructions

Key Novelty.

- Introduce a *constraint adherence score* beyond standard text similarity metrics
- Penalize violations of time, calorie, or diet constraints

Why This Fits the Dataset. The dataset explicitly encodes constraints, making it possible to evaluate whether generated instructions respect them.

Idea 2: Recipe Instruction Simplification

Problem. Recipe instructions are often verbose, contain redundant steps, and are not beginner-friendly.

Research Question. Can cooking instructions be automatically simplified without losing procedural correctness?

Task Definition.

- **Input:** Original recipe instructions
- **Output:** Simplified instructions with fewer steps

Why This Fits the Dataset.

- Instructions are procedural and step-based
- Many steps repeat similar actions (e.g., stirring, heating)

Evaluation Signals.

- Compression ratio
- Semantic similarity to original instructions
- Preservation of cooking order and time consistency

Idea 3: Ingredient–Action Alignment

Problem. Text generation models often perform actions without explicitly grounding them in ingredients.

Research Question. Can we learn fine-grained alignment between ingredients and cooking actions?

Task Definition. Extract structured tuples from instructions, such as:

$$(onion \rightarrow chop), \quad (beef \rightarrow sauté), \quad (flour \rightarrow whisk)$$

Method.

- Dependency parsing and sequence labeling
- Weak supervision using ingredient lists

Output.

- Ingredient–action graphs
- Action coverage scores per recipe

Strength. This is an information extraction task rather than text generation, leading to higher precision and lower hallucination.

Idea 4: Hallucination Detection in Recipe Generation

Problem. Generated recipes often introduce ingredients or steps that are not supported by the input.

Research Question. How can hallucinated ingredients or steps be detected in generated cooking instructions?

Task Definition.

- **Input:** Ingredient list + instruction step
- **Output:** Grounded or Hallucinated

Automatic Labeling Strategy.

- Ingredient present in list → valid
- Ingredient absent from list → hallucinated

Novelty. Applies hallucination detection beyond question answering, focusing instead on procedural instructional text.

Conclusion

These four ideas demonstrate how a structured recipe dataset can support diverse and precise NLP research problems, including constraint-aware generation, simplification, information extraction, and hallucination detection. Each idea can be implemented independently or combined into a larger project on trustworthy and grounded language generation.