Identifying Key Entities in Recipe Data-Report

Submitted by: Elakkiya Chezhiyan

Submitted to: UpGrad - PG Program in Machine Learning & AI

Executive Summary

This project focuses on identifying key entities, ingredients, quantities, and units, from unstructured recipe text using a sequence labeling approach. A Conditional Random Field (CRF) model was used to classify each token in a recipe sentence. The assignment involved preprocessing recipe data, engineering features, balancing classes, training a CRF model, and performing error analysis. The model performed well on structured patterns and offered insights into areas for improvement. The output supports applications like recipe parsers, grocery planners, or virtual kitchen assistants.

Problem Statement

Recipes contain structured information presented in unstructured text. Extracting the key components, ingredient names, measurement units, and quantities, is essential for many applications. This assignment formulates it as a Named Entity Recognition (NER) task, where each word (token) in a recipe is classified into one of the three categories.

Methodology

1. Data Loading and Validation:

- Loaded recipe data from a JSON file containing token-level POS labels.
- Cleaned and validated the data by ensuring that input_tokens and pos_tokens were aligned in length and structure.
- Dropped inconsistent rows and recalculated token lengths for reliability.

	input	pos	input_tokens	pos_tokens
0	6 Karela Bitter Gourd Pavakkai Salt 1 Onion 3 tablespoon Gram flour besan 2 teaspoons Turmeric powder Haldi Red Chilli Cumin seeds Jeera Coriander Powder Dhania Amchur Dry Mango Sunflower Oil	quantity ingredient ingredient ingredient ingredient ingredient quantity ingredient quantity unit ingredient ingredient ingredient quantity unit ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient	[6, Karela, Bitter, Gourd, Pavakkai, Salt, 1, Onion, 3, tablespoon, Gram, flour, besan, 2, teaspoons, Turmeric, powder, Haldi, Red, Chilli, Cumin, seeds, Jeera, Coriander, Powder, Dhania, Amchur, Dry, Mango, Sunflower, Oil]	[quantity, ingredient, ingredient, ingredient, ingredient, ingredient, quantity, ingredient, quantity, unit, ingredient, ingredient, ingredient, quantity, unit, ingredient, ingredient, ingredient, ingredient, ingredient, ingredient, ingredient, ingredient, ingredient, ingredient, ingredient, ingredient, ingredient, ingredient, ingredient, ingredient]
1	2-1/2 cups rice cooked 3 tomatoes teaspoons BC Belle Bhat powder 1 teaspoon chickpea lentils 1/2 cumin seeds white urad dal mustard green chilli dry red 2 cashew or peanuts 1-1/2 tablespoon oil asafoetida	quantity unit ingredient ingredient quantity ingredient unit ingredient ingredient ingredient ingredient quantity unit ingredient ingredient quantity ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient quantity ingredient ingredient ingredient quantity unit ingredient ingredient	[2-1/2, cups, rice, cooked, 3, tomatoes, teaspoons, BC, Belle, Bhat, powder, 1, teaspoon, chickpea, lentils, 1/2, cumin, seeds, white, urad, dal, mustard, green, chilli, dry, red, 2, cashew, or, peanuts, 1-1/2, tablespoon, oil, asafoetida]	[quantity, unit, ingredient, ingredient, quantity, ingredient, unit, ingredient, ingredient, ingredient, ingredient, quantity, unit, ingredient, ingredient, quantity, ingredient, ingredient, ingredient, ingredient, ingredient, ingredient, ingredient, ingredient, ingredient, ingredient, quantity, ingredient, ingredient, ingredient, quantity, unit, ingredient, ingredient]
2	1-1/2 cups Rice Vermicelli Noodles Thin 1 Onion sliced 1/2 cup Carrots Gajjar chopped 1/3 Green peas Matar 2 Chillies 1/4 teaspoon Asafoetida hing Mustard seeds White Urad Dal Split Ghee sprig Curry leaves Salt Lemon juice	quantity unit ingredient ingredient ingredient ingredient quantity ingredient ingredient quantity unit ingredient ingredient ingredient quantity ingredient ingredient ingredient quantity ingredient quantity unit ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient unit ingredient ingredient ingredient ingredient ingredient	[1-1/2, cups, Rice, Vermicelli, Noodles, Thin, 1, Onion, sliced, 1/2, cup, Carrots, Gajjar, chopped, 1/3, Green, peas, Matar, 2, Chillies, 1/4, teaspoon, Asafoetida, hing, Mustard, seeds, White, Urad, Dal, Split, Ghee, sprig, Curry, leaves, Salt, Lemon, juice]	[quantity, unit, ingredient, ingredient, ingredient, ingredient, quantity, ingredient, ingredient, quantity, unit, ingredient, ingredient, ingredient, quantity, ingredient, ingredient, ingredient, quantity, ingredient, quantity, unit, ingredient, ingredient, ingredient, ingredient, ingredient, ingredient, ingredient, ingredient, ingredient, unit, ingredient, ingredient, ingredient, ingredient, ingredient]
3	500 grams Chicken 2 Onion chopped 1 Tomato 4 Green Chillies slit inch Ginger finely 6 cloves Garlic 1/2 teaspoon Turmeric powder Haldi Garam masala tablespoon Sesame Gingelly Oil 1/4 Methi Seeds Fenugreek Coriander Dhania Dry Red Fennel seeds Saunf cups Sorrel Leaves Gongura picked and	quantity unit ingredient quantity ingredient ingredient quantity ingredient quantity ingredient ingredient ingredient unit ingredient ingredient quantity unit ingredient quantity unit ingredient ingredient ingredient ingredient ingredient unit ingredient ingredient ingredient quantity ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient unit ingredient ingredient ingredient ingredient ingredient	[500, grams, Chicken, 2, Onion, chopped, 1, Tomato, 4, Green, Chillies, slit, inch, Ginger, finely, 6, cloves, Garlic, 1/2, teaspoon, Turmeric, powder, Haldi, Garam, masala, tablespoon, Sesame, Gingelly, Oil, 1/4, Methi, Seeds, Fenugreek, Coriander, Dhania, Dry, Red, Fennel, seeds, Saunf, cups, Sorrel, Leaves, Gongura, picked, and]	[quantity, unit, ingredient, quantity, ingredient, ingredient, quantity, ingredient, quantity, ingredient, ingredient, ingredient, unit, ingredient, ingredient, quantity, unit, ingredient, quantity, unit, ingredient, ingredient, ingredient, ingredient, ingredient, unit, ingredient, ingredient, ingredient, quantity, ingredient, ingredient, ingredient, ingredient, ingredient, ingredient, ingredient, ingredient, ingredient, unit, ingredient, ingredient, ingredient, ingredient, ingredient]
4	1 tablespoon chana dal white urad 2 red chillies coriander seeds 3 inches ginger onion tomato Teaspoon mustard asafoetida sprig curry	quantity unit ingredient ingredient ingredient ingredient quantity ingredient ingredient ingredient ingredient quantity unit ingredient ingredient ingredient unit ingredient ingredient unit ingredient	[1, tablespoon, chana, dal, white, urad, 2, red, chillies, coriander, seeds, 3, inches, ginger, onion, tomato, Teaspoon, mustard, asafoetida, sprig, curry]	[quantity, unit, ingredient, ingredient, ingredient, ingredient, quantity, ingredient, ingredient, ingredient, ingredient, quantity, unit, ingredient, ingredient, ingredient, unit, ingredient, ingredient, unit, ingredient]

2. Exploratory Data Analysis (EDA):

- Visualized the most frequent ingredients and units using bar plots.
- Analyzed both training and validation datasets to identify distribution imbalances and label trends.

```
Top 10 most frequent ingredients in Training dataset:
powder: 129
Salt: 102
seeds: 89
Green: 85
chopped: 84
Oil: 83
Red: 81
Chilli: 77
Coriander: 71
Sunflower: 65

Top 10 most frequent units in Training dataset:
teaspoon: 162
cup: 136
tablespoon: 99
grams: 63
tablespoons: 61
inch: 52
cups: 50
sprig: 41
cloves: 39
teaspoons: 39
```

3. Feature Engineering:

- Designed a word2features function using spaCy for token-level features including:
  - Lexical (token, lemma, shape)
  - POS and dependency tags
  - Digit and punctuation flags
  - Quantity/unit detection using regex and keyword sets
  - Contextual features (previous and next tokens)

```
# print the length of train features and labels
print("Length of X_train_features:", len(X_train_features))
print("Length of y_train_labels:", len(y_train_labels))
```

```
Length of X_train_features: 196
Length of y_train_labels: 196
```

```
# print the length of validation features and labels
print("Length of X_val_features:", len(X_val_features))
print("Length of y_val_labels:", len(y_val_labels))
```

```
Length of X_val_features: 84
Length of y_val_labels: 84
```

4. Class Weighting:

- Used inverse frequency to compute class weights, with additional penalization on 'ingredient' to address misclassification risk.

```
Label Counts: Counter({'ingredient': 5323, 'quantity': 980, 'unit': 811})
Total Samples: 7114
```

5. Model Building:

- Trained a CRF model with hyperparameters:
- algorithm='lbfgs', c1=0.5, c2=1.0, max_iterations=100, all_possible_transitions=True
- Used token-level feature dictionaries with embedded class weights during training.

	precision	recall	f1-score	support
ingredient	1.00	1.00	1.00	5323
quantity	0.99	0.98	0.99	980
unit	0.98	0.99	0.98	811
accuracy			1.00	7114
macro avg	0.99	0.99	0.99	7114
weighted avg	1.00	1.00	1.00	7114

Visualizations and Key Insights

1. Training Performance:

- High accuracy for 'unit' and 'quantity' labels.
- Most misclassifications occurred in the 'ingredient' class due to semantic overlap and varied context.

2. Validation Metrics:

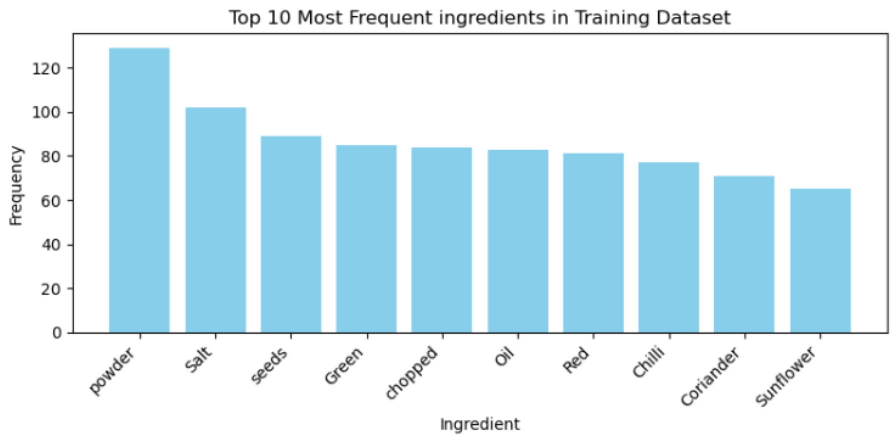
- Flat classification report showed precision and recall scores across labels.
- Confusion matrix highlighted label confusions, particularly between 'ingredient' and 'unit'.

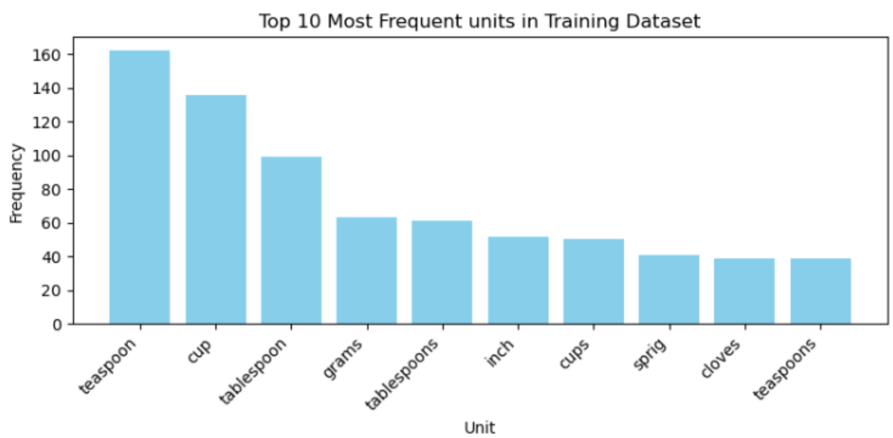
3. Error Analysis:

- Errors were logged when predicted labels did not match the true labels.
- Tokens like 'oil', 'milk', and 'salt' were often misclassified due to ambiguous context.
- Accuracy dropped when quantities were missing or sentence context was vague.
- A detailed error table was created with token, true label, predicted label, and surrounding context.

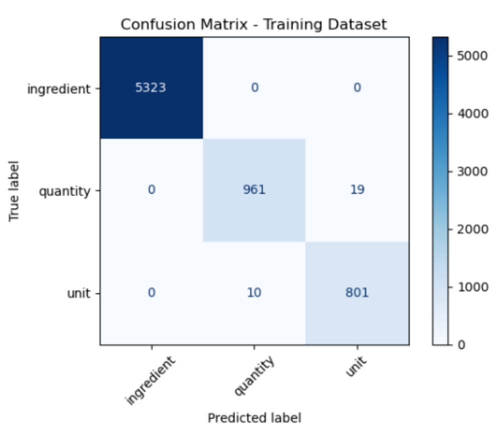
4. Sample Visualizations Included:

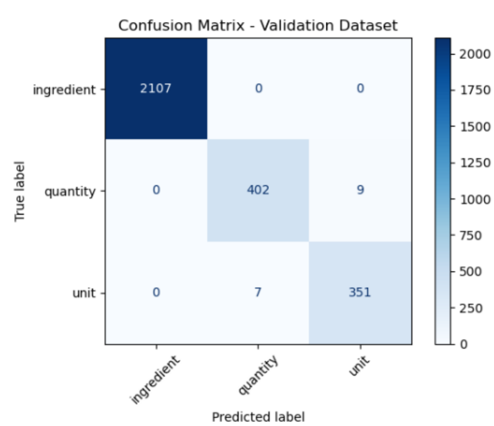
- Bar charts for top 10 ingredients and units.





- Confusion matrices for both training and validation datasets.





➢ Confusion matrix training data.

➢ The table below shows error analysis with label-wise metrics and sample misclassifications from the validation set

Per-Label Error Analysis:

	label	total	errors	accuracy	class_weight
1	quantity	411	9	97.81	7.26
2	unit	358	7	98.04	8.77
0	ingredient	2107	0	100.00	3.01

	token	prev_token	next_token	true_label	pred_label	context
0	1/4	powder	Salt	quantity	unit	powder 1/4 Salt
1	cut	French	into	unit	quantity	French cut into
2	6-8	Pudina	Saffron	quantity	unit	Pudina 6-8 Saffron
3	cold	Oil	pressed	unit	quantity	Oil cold pressed
4	1-1/2	pressed	Poppy	quantity	unit	pressed 1-1/2 Poppy
5	into	cut	cm	unit	quantity	cut into cm
6	2	1-1	tablespoon	quantity	unit	1-1 2 tablespoon
7	1/3	powder	Water	quantity	unit	powder 1/3 Water
8	pinch	powder	Salt	unit	quantity	powder pinch Salt
9	2	Salt	Instant	quantity	unit	Salt 2 Instant

Assumptions

- The `input` and `pos` sequences are pre-aligned and consistently tokenized.
- Quantity detection assumes common English fraction and decimal formats.
- Unit keywords are manually defined and may not cover all variations (e.g., 'dash', 'pinch').

Value & Benefits

- ✓ **Enhanced User Experience**: Automatically structuring recipe text enables voice assistants and mobile apps to guide users step-by-step, interpret ingredient lists, or convert units on the fly.
- ✓ **Smart Grocery Planning**: By extracting ingredients and quantities, apps can generate dynamic shopping lists, check pantry stock, or suggest meal prep based on available items.
- ✓ **Data Integration**: Structuring recipes allows integration with **nutrition APIs**, **meal tracking apps**, or **smart kitchen devices**, adding value across fitness, healthcare, and lifestyle sectors.
- ✓ **E-commerce Enablement**: Online retailers can link extracted ingredients to purchasable items, driving personalized product recommendations and increasing cart conversions.
- ✓ **Scalable Automation**: The CRF-based model is lightweight, interpretable, and deployable in resource-constrained environments like edge devices or mobile platforms.

Conclusion

The project achieved its objective of extracting structured information from recipe text using a CRF model with carefully engineered features and class weighting. The model performed strongly on consistent patterns like numeric quantities and units, and revealed areas for further improvement in handling context-heavy ingredient names. The work establishes a solid baseline for expanding this pipeline to real-world cooking assistants, shopping list generators, or voice-driven recipe platforms.