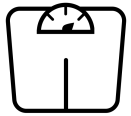
FOOD LOG TAGGER

Stephanie Chong

About



COMPANY

weight management program + health data tracking & monitoring + clinical guidance

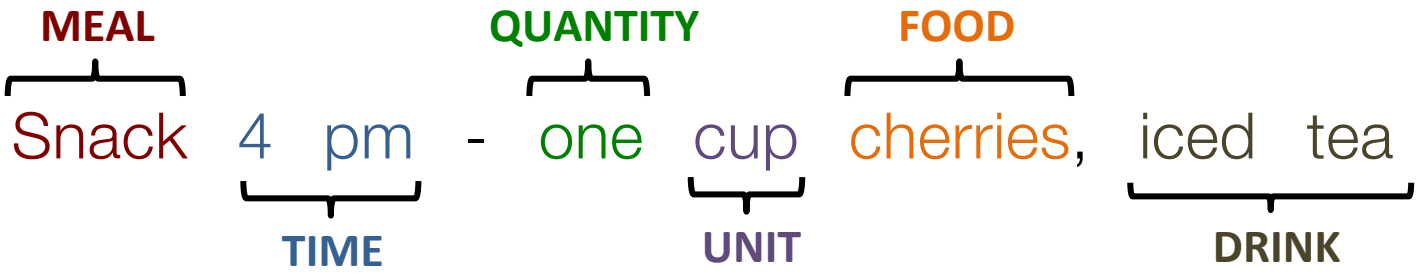
PROBLEM

customer-generated nutrition data exists in unstructured format

GOAL

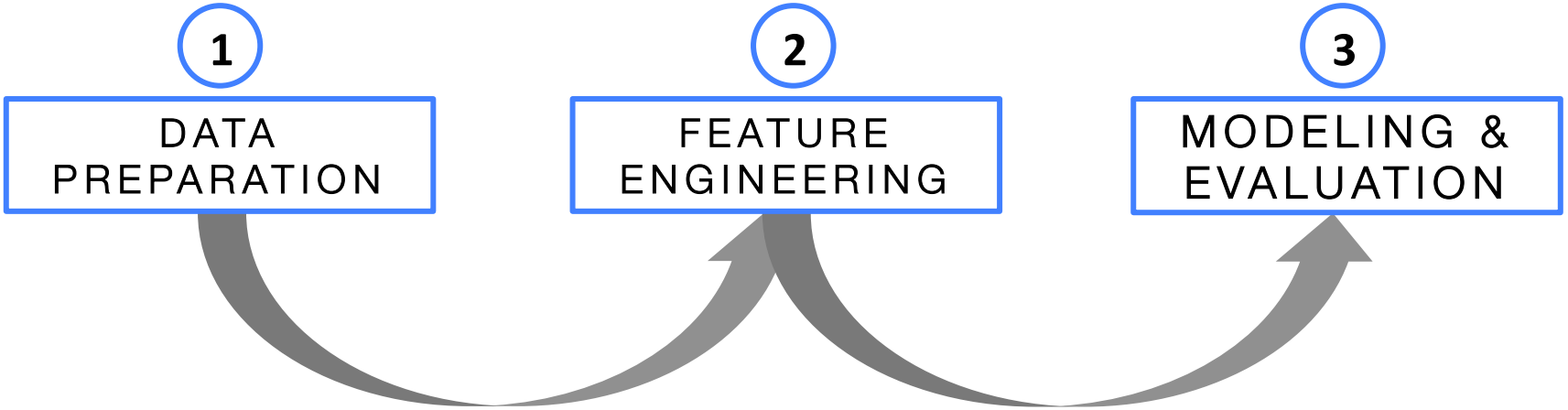
predict label for each word in a food log

Data


MEAL
QUANTITY
FOOD
Snack 4 pm - one cup cherries, iced tea
TIME
UNIT
DRINK


COMMENT
Last night I was really low on my calorie intake so I treated myself.
Had brie, walnuts, and dried peaches.
OTHER
FOOD
FOOD
OTHER
FOOD

Process


1
DATA PREPARATION
2
FEATURE ENGINEERING
3
MODELING & EVALUATION

Process

1. **DATA PREPARATION**
   - Manually label training data using Flask web app + MongoDB
   - Labels:
     1. Meal
     2. Time
     3. Food
     4. Drink
     5. Quantity
     6. Unit
     7. Comment
     8. Other
2. **FEATURE ENGINEERING**
3. **MODELING & EVALUATION**

Process

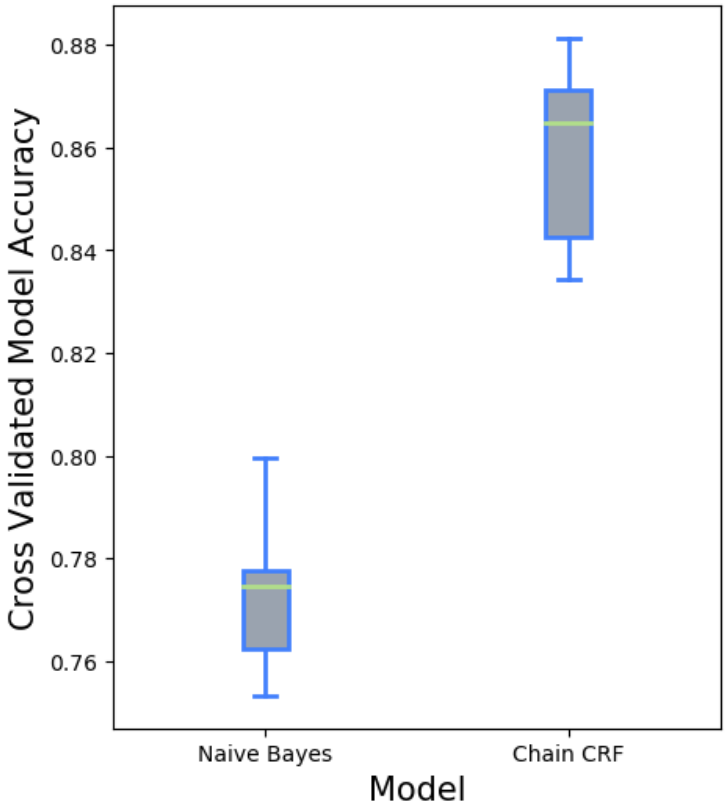
1. DATA PREPARATION
2. FEATURE ENGINEERING
   - spaCy part-of-speech tags
   - Boolean: word has numeric digit
   - Boolean: word has punctuation
   - Boolean: word is capitalized
   - POS tags for words before and after
3. MODELING & EVALUATION

Process

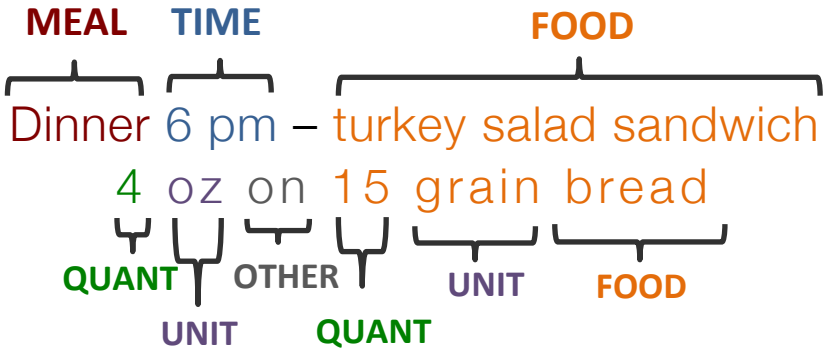
1. DATA PREPARATION
2. FEATURE ENGINEERING
3. MODELING & EVALUATION

- Linear Chain Conditional Random Fields (CRF) → graphical model that maximizes p(y|x)
- Baseline: Naïve Bayes
- Performance metric: accuracy

Results



Chain CRF Prediction:



Next Steps

Optimize model given more data

☺☺☹☹

Topic Modeling / Sentiment Analysis on ‘comments’

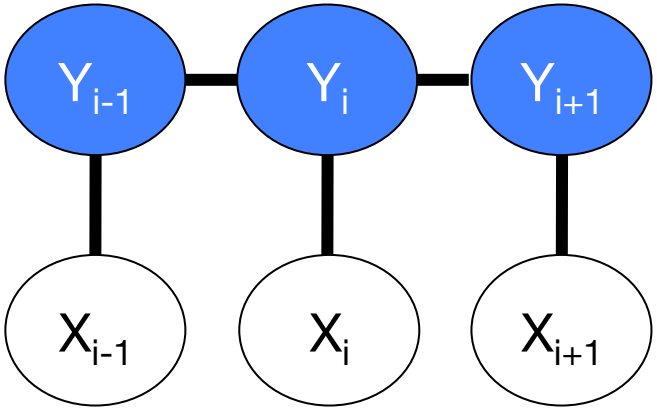
Regression on weight outcomes

Thank You!


github.com/ssychong


linkedin.com/in/ssychong

Appendix: CRF



CRF model consists of vertices, edges, and unary & pair-wise potentials.
X represents elements of the structured input (in this case, words); Y represents the labels for each input (in this case, tags).

IMPLEMENTATION

Given a sequence $X = [x_1, x_2, \ldots, x_n]$, predict the corresponding labels $Y = [y_1, y_2, \ldots, y_n]$

1. Creation of feature vectors from the given sequence. $F = [f_1, f_2, \ldots, f_n]$
2. Calculate probability of each possible label y_i given a feature vector f_i
3. The y_i with maximum probability for given feature f_i is the label of token x_i. Label of x_i = argmax($P(y_i|f_i)$)

Appendix: Per-Class Scoring

LABEL	PRECISION	RECALL	F1-SCORE	SUPPORT
meal	0.85	0.84	0.84	789
time	0.79	0.60	0.68	208
food	0.87	0.97	0.92	2604
drink	1.00	0.01	0.02	96
quantity	0.89	0.85	0.87	219
unit	0.81	0.24	0.37	107
comment	0.94	0.88	0.91	964
other	0.88	0.85	0.87	258
avg / total	0.88	0.88	0.86	5245