Food Recipe Processing & Entity Recognition

Kevin Pang & Kathryn Hamilton

W266 Summer 2018

Problem Introduction

Test-Based Recipe → *Entity Recognition* → *Relationship Extraction*

Grandma's Stuffed Zucchini Edit

Ingredients

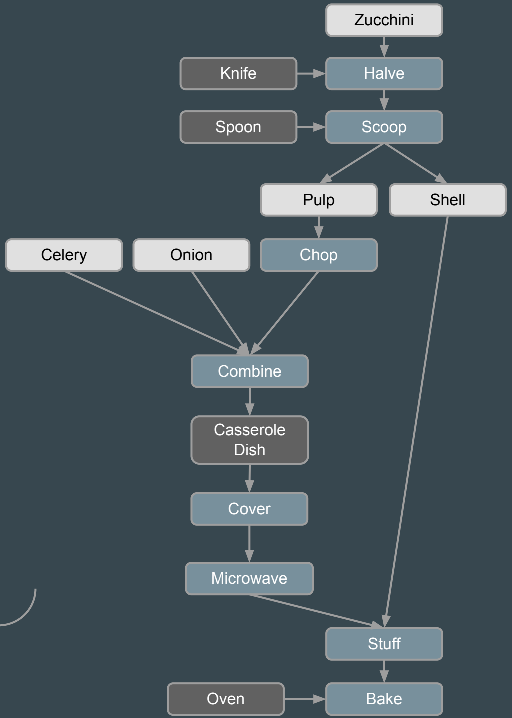
- 2 Large Zucchini
- 1 Small Onion (Chopped)
- 1 Cup Mushrooms (Coarsly chopped)
- 1/2 Cup Celery (Chopped)
- 1 Medium Tomato (Chopped)
- 1 Tbsp Butter

Directions

1. Halve zucchini lengthwise. Scoop out pulp and set aside, leaving 1/4" shell.
2. Chop pulp coarsely.
3. Combine with onion, mushrooms, celery, and tomatoes in 2 quart casserole dish. Cover. Microwave on high for 4-6 minutes or until tender. Stir once and drain.

'Combine with onion, mushrooms, celery, and tomatoes in 2 quart casserole dish. Cover. Microwave on high for 4-6 minutes or until tender. Stir once and drain.'

'Combine Activity with onion Vegetable, mushrooms Vegetable, celery Vegetable, and tomatoes Vegetable in 2 quart casserole dish Tool. Cover Activity. Microwave Activity on high for 4-6 minutes or until tender. Stir Activity once and drain Activity.'

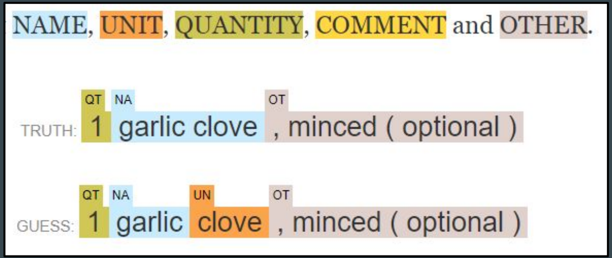


Our Scope

Previous Work

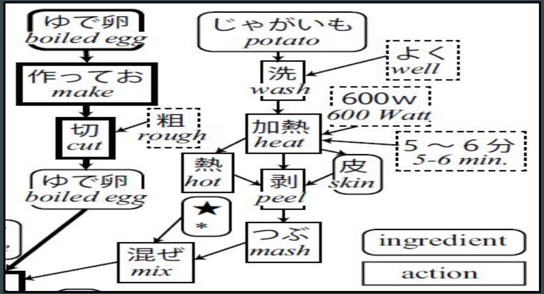
NER: *Partially Solved*

- **Open-ended**
- High accuracy **only in certain structured scenarios**



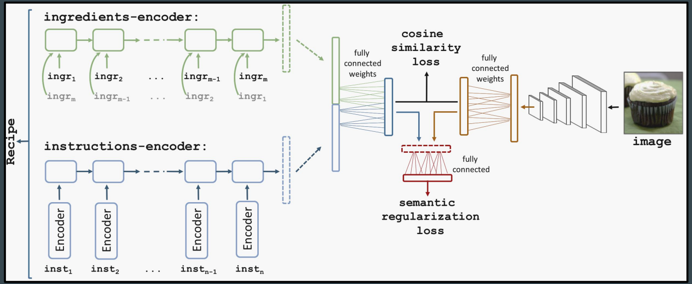
Relationship Extraction: *Unsolved*

- **No universal DAG format exists**
- **Fairly inaccurate** (~60%)
- Difficulties: nonatomic phrases, coreference resolution, dropped words



Other Recipe-Related Work:

- **Generate recipe from image** and vice versa with 65% accuracy



Our Framework

Dataset

Recipe 1M

One million recipe and image pairs. Has not been used in the context of DAGs. Does not have any DAG-related labels.

Ingredients:

{'text': '1/2 pound lean beef, preferably sirloin, sliced as thinly as possible'}

Instructions:

{'text': 'Mix the beef with the garlic, soy sauce, and sesame oil and marinate for a few minutes.'}

NLP

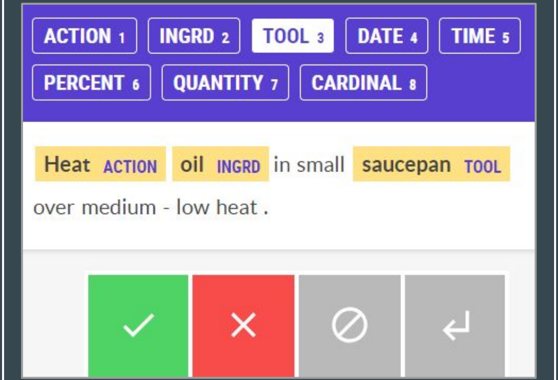
spaCy

Open-source, industrial strength library for advanced NLP in Python.

Ground Truth

Prodigy (Active Learning)

Machine teaching API and UI to efficiently train and evaluate models. Used to create 350 ground truth examples.



Our Approach

Seeding

Using seeding words scraped from the internet (actions, tools, ingredients), extend a prebuilt spaCy NER model.

Compare predicted entities against ~350 manually annotated instructions.

Precision	Recall	F1-Score
0.814	0.524	0.638

Low recall → need training data to be more representative and capture more patterns

Bootstrapping

Augment the seeding words set using a variety of techniques:

- Use patterns instead of words
  (eg. whisking, whisked → whisk)
- Use a custom embedding
- Use machine learning to learn more patterns
  (Prodigy's ML Interface)

Precision	Recall	F1-Score
0.770	0.577	0.660

Embedding focused on single words; multi-word entities are limited by complexity of pattern matching

Neural Network

Use bootstrapped seeding words and patterns to efficiently annotate more ground truth instructions.

Experiment with different NN models.

Eg. Train a CNN (depth 4) with ~450 ground truth instructions, 70/30 split.

Precision	Recall	F1-Score
0.850	0.856	0.853

Using bs = 15 , iters = 10, dropout = 0.2

Conclusions & Further Work

Training generalized models on **unbounded** and **evolving** problems is **very difficult**.

Working with unlabelled data is difficult. Trying to create labelled data for **unstructured, highly-variable text** is also very difficult.

However, **active learning**, **machine learning**, and **deep learning** can **collaboratively improve** the model based on a feedback loop, which is also a more intuitive process compared to a black box.

In a short time, we **improved recall from 0.524 to 0.856 in NER tasks.**

We believe the **same approach can be extended to relationship extraction** and structural representation.

By combining named entities and relationships, we can move one step closer to an effective directed acyclic graph (DAG) representation of instructional text.

Questions?

Kevin Pang

Kevin is an analytics practitioner in the insurance industry with over a decade of experience in providing data-driven impact.

Kathryn Hamilton

Kathryn is a systems engineer and mathematician in the automotive industry with a focus on autonomy and mobility services.