

Popular Dishes

A Deep Dive into an ML Project at Yelp

Matt Smith
matts@yelp.com



About Me

Software engineer @ Yelp on the
Semantic Business Info team

2 years Mechatronics Engineering ->
1 year Software Engineering ->
2 years CS + C&O

♥ board games, cooking, language
learning, math



My Path into Data Science

1st coop as a cyber-security developer

2nd/3rd coop: Yelp operations intern



Autoscaling PaaS Services



Matt S., Software Engineer

May 25, 2016

If you haven't heard about PaaS before, feel free to check out the b

One step in creating a service is to decide how many compute resour
inception of PaaS, changing a service's resource allocation has req
pushing new configs, and service authors had to pore over graphs and
proper resource allocation for a service whenever load requirements c
earlier this month when autoscaling was introduced into PaaS.

My Path into Data Science



Can I do data science now?

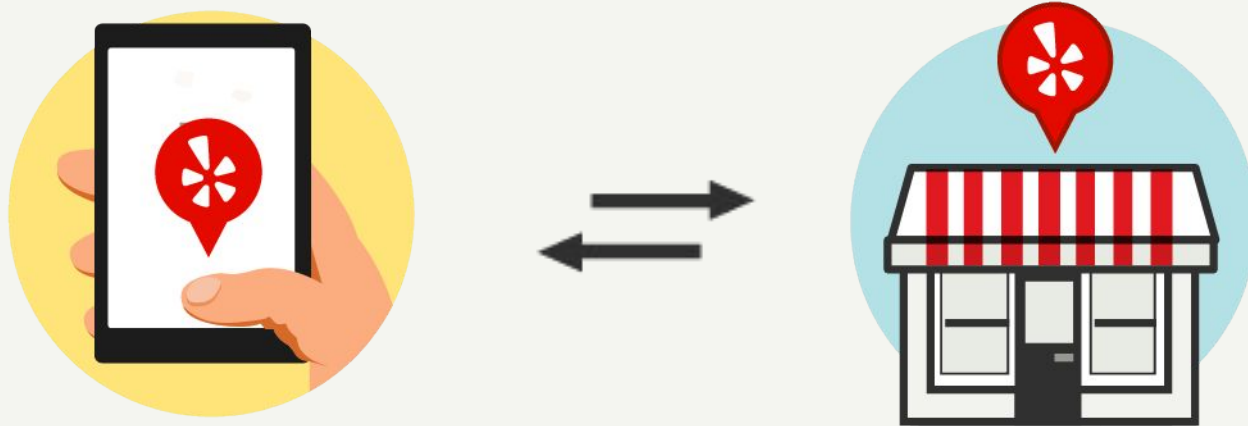
(I did Andrew Ng's ML course)

Sure!



Yelp's Mission

Connecting people with great local businesses.



What is Popular Dishes?



Popular Dishes **Mission**

Ease cognitive burden

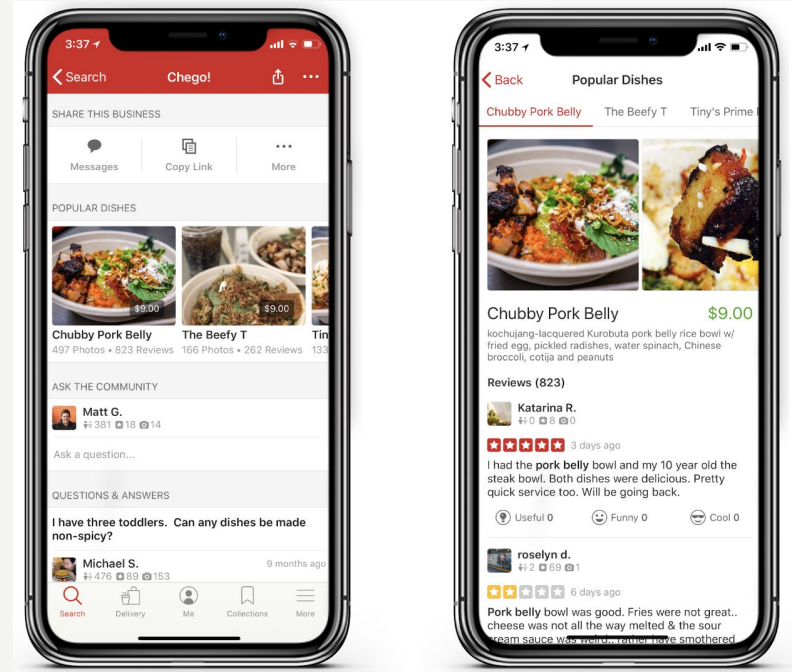
Show users the best menu items

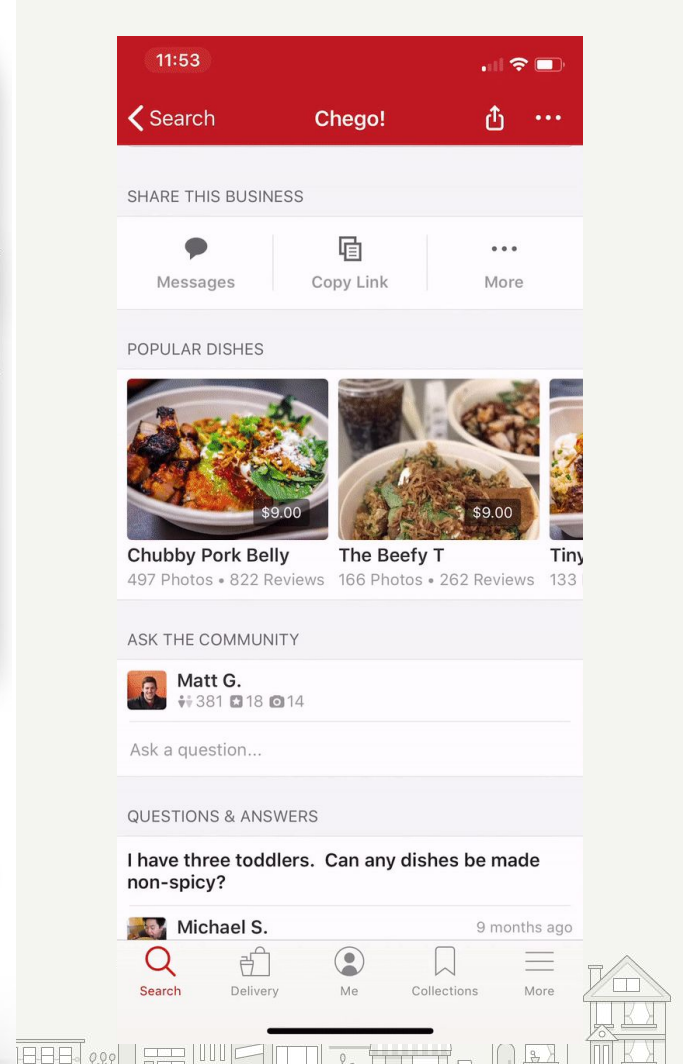
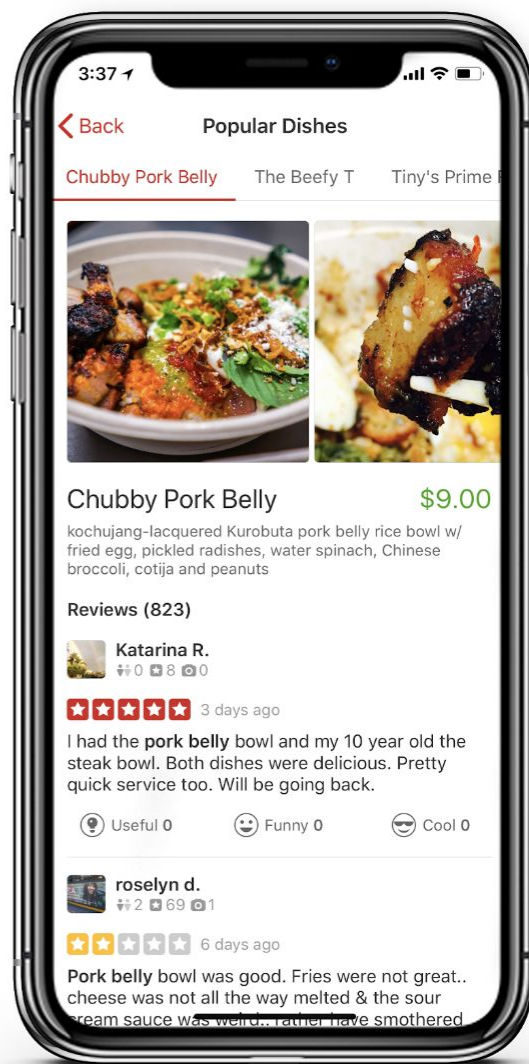
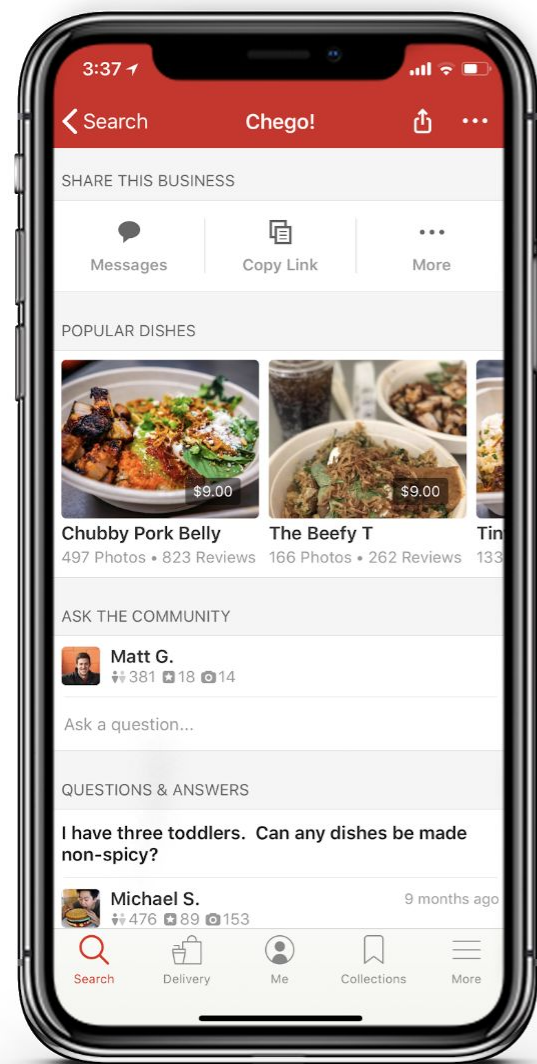
Aid in discovery

“Take the guesswork out of what to order”

Beautify

Make the existing UI more beautiful





Popular Dishes **Inspiration**

Highlights service

Show popular menu items in reviews

Already on the biz page, just needs a UI update



"While my **risotto** and quail were masterpieces in their own right, I think they over salted my dishes." in 959 reviews

[Risotto](#)



"My personal favorites are foie gras, oyster with caviar, lobster salad, **duck breast**, macaroon ice cream sandwich, and their cheese cart." in 323 reviews

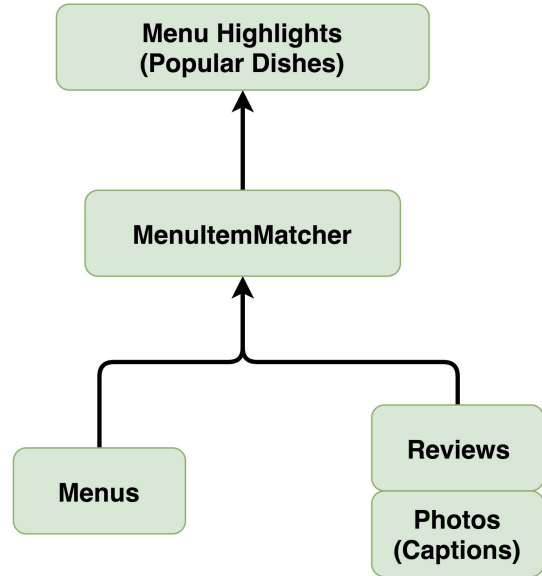
[Lemon Pepper Duck Breast](#)



"- **Roast Maine Lobster** with Mushrooms, Corn and Tarragon -- Yes, more lobster, and I LOVE IT!!!" in 414 reviews

[Roast Maine Lobster](#)

[Show more review highlights](#)

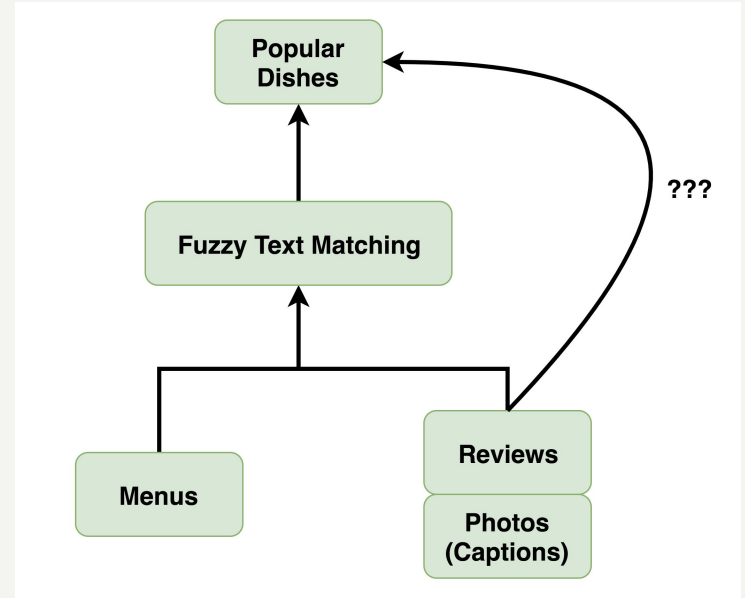


Popular Dishes **Problem**

Only 30% of restaurants have menus

Users won't open Yelp to check Popular Dishes if it's only there 30% of the time

What would a solution look like?



Non-ML Ideas

Sometimes we think of ML as a solution to any problem

But for simple problems, ML

- Takes longer
- Creates more tech debt
- Is less efficient



**Machine Learning:
The High-Interest Credit Card of Technical Debt**

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov,
Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young
{dsculley, gholt, dgg, edavydov}@google.com
{toddpillips, ebner, vchaudhary, mwyoung}@google.com
Google, Inc

Non-ML Ideas

Idea: Get more menus

- We could try to encourage businesses to submit menus
- Pay a company to scrape menus for us
- Do OCR on user-submitted menu photos
- Scrape business web-pages for menu data

Ultimately this won't get us enough menus

The last two ideas are also very difficult!




Non-ML Ideas

Idea: If a restaurant sells a dish it's probably on a menu somewhere

Create a universal menu merging all of our menus into one

Problem: Fuzzy matching is slow on big menus, no signature dishes

Popular Dishes



Pork Sisig Burrito
25 Reviews • 5 Photos

Spicy Senior Burrito
31 Reviews • 2 Photos

The image shows two burritos side-by-side. The one on the left is a Pork Sisig Burrito, filled with a chunky, saucy meat mixture (sisig) and topped with green onions. The one on the right is a Spicy Senior Burrito, filled with a chunky, saucy meat mixture (senior) and topped with green onions. Both burritos are wrapped in white flour tortillas and are shown in a close-up shot.

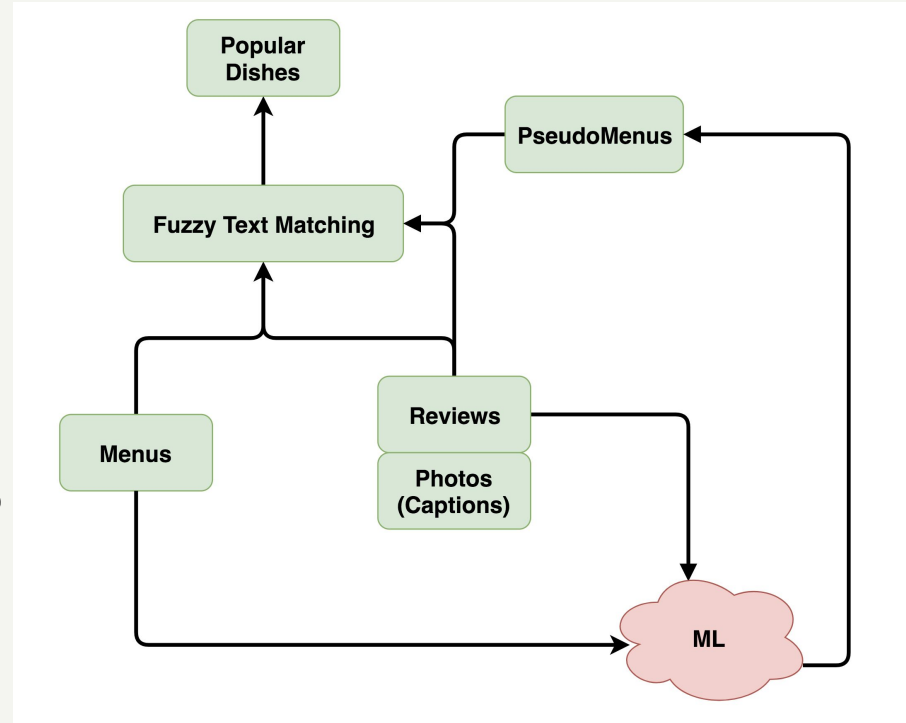


ML Ideas

Alright... you've convinced me.

We'll use ML to bridge the gap between reviews and Popular Dishes.

But how to implement this ML?



ML Cookbook

You need:

- *A way of turning data -> training data*
- A dataset
- A model
- Some infrastructure
- Evaluation metrics



ML Training Data

Having an algorithm read reviews for a business and output its entire menu is hard

Litmus test for feasibility: can **you** do that?

Recommended Reviews for Gary Danko

Search within the reviews



Sort by **Yelp Sort** ▾

Language **English (4978)** ▾

Please read all 5000 reviews and then tell me the menu from memory



ML Training Data

What I can do:

What things in a review are food?

“I got the **spicy chicken burrito**
and the **nachos**”

Do this for each review and
merge the results



ML Training Data

This is a well-known sequence classification problem

- Sequence labelling
 - Mark sections in sequences with labels
 - Named Entity Recognition (NER)
 - John Doe lives in Canada ->
(John Doe)(Person) lives in (Canada)(Place)
 - Food Recognition
 - The risotto was amazing ->
The (risotto)(Food) was amazing



ML Training Data

We don't just mark things as food.

- BIESO Labels
 - **B**egin, **I**nside, **E**nd, **S**ingleton, **O**utside
 - (The **O**) (Spicy **B**) (Chicken **I**) (Burrito **E**) (and **O**) (nachos **S**) (were **O**) (delicious **O**)
 - Some sequences can be invalid (eg: **O,O,B,O,E**)

We do this to differentiate adjacent words:

“I got the **steak, risotto** and **greek salad**” vs

“I got the (steak **S**), (risotto **S**) and (greek **B**) (salad **E**)”



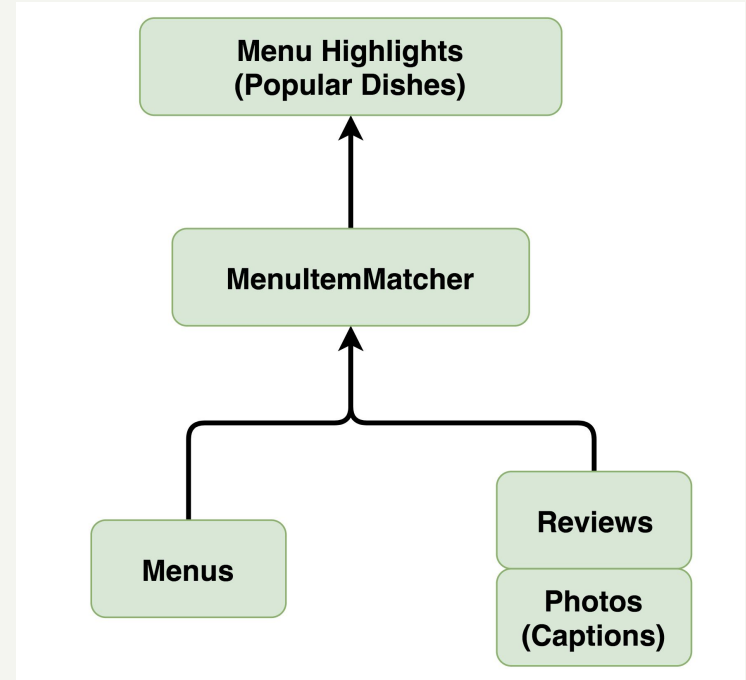
ML Training Data

Okay but...

To train a model to do this tagging we need some training examples

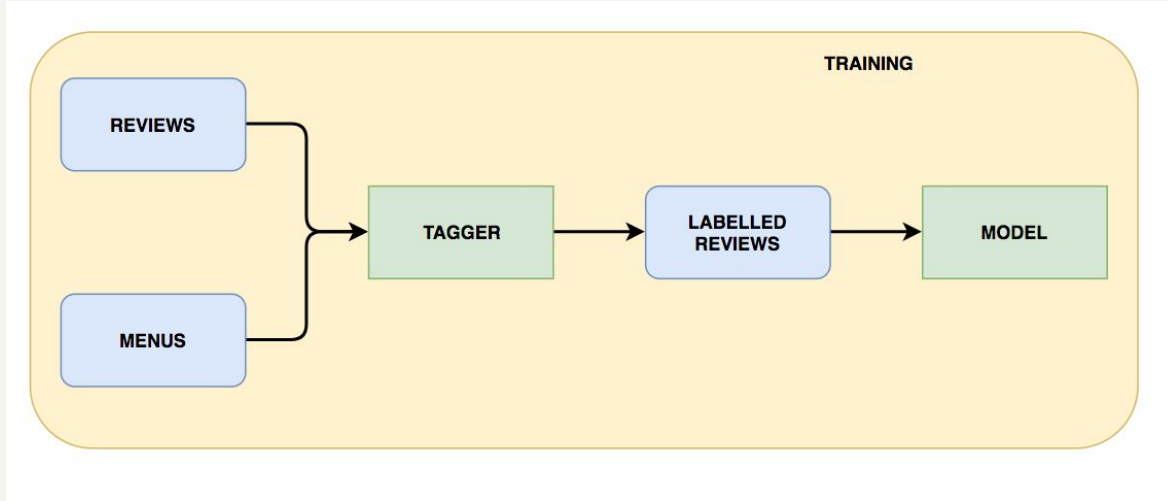
Our existing tools look promising

We can modify our existing matcher to tag the matches with BIESO labels



ML Training Data

The result:



We've created a way to turn our data into training data!



ML Cookbook

You need:

- A way of turning data -> training data
- **A dataset**
- A model
- Some infrastructure
- Evaluation metrics



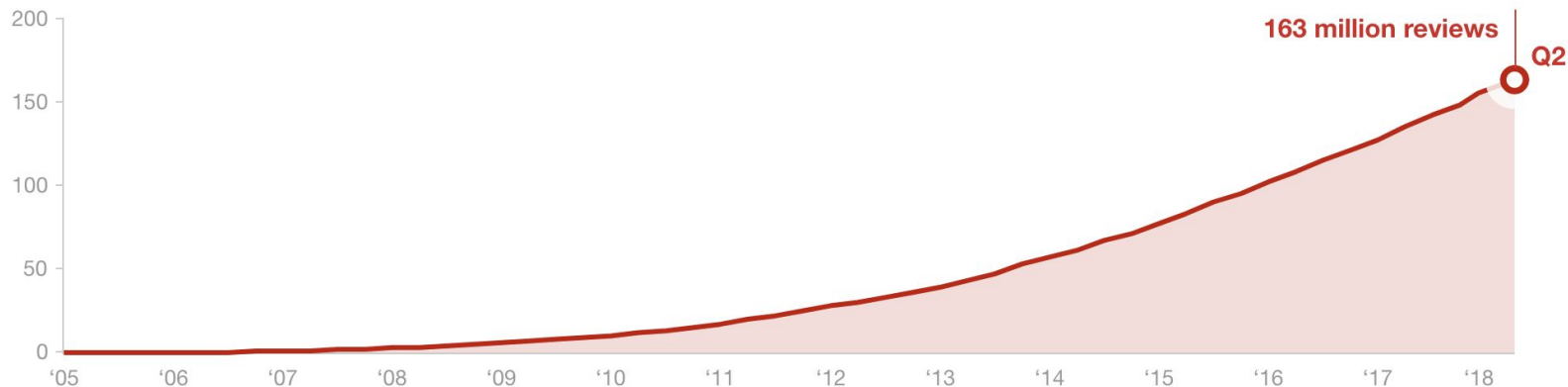
ML Dataset

We'll just make a naive choice for now:

Restaurants with menus and at least 5 reviews

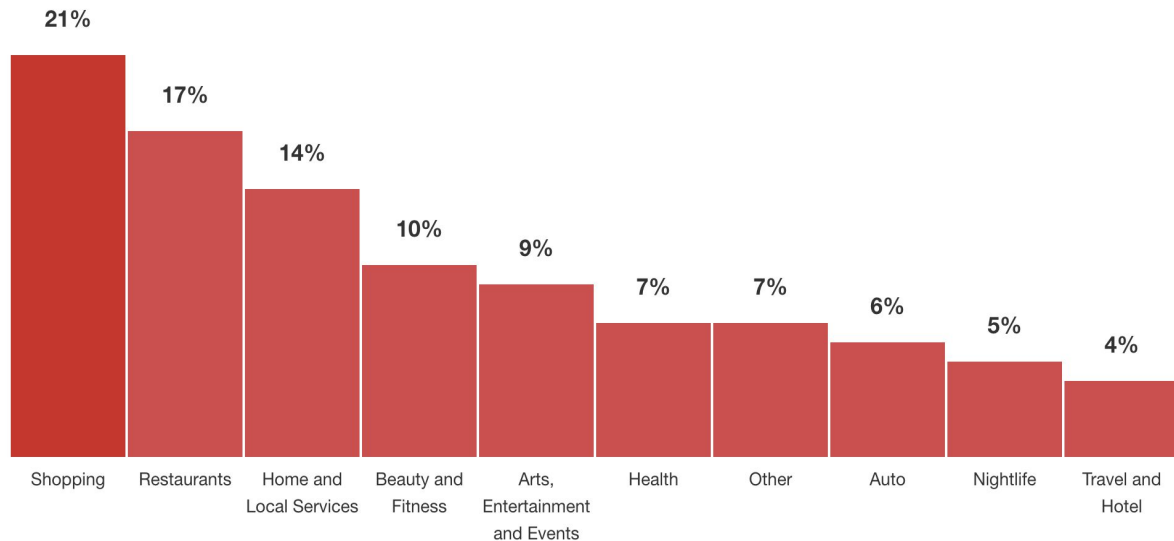
Review Statistics as of June 30, 2018

Cumulative reviews contributed since inception

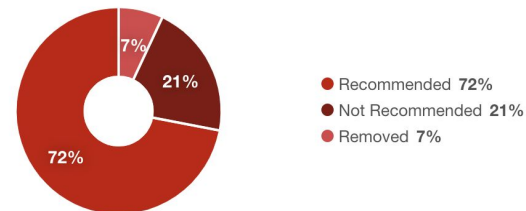


ML Dataset

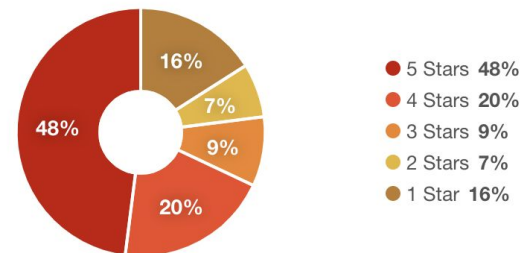
Reviewed Businesses by Category



Recommended Distribution



Rating Distribution



~20 million restaurant reviews
30% of restaurants have menus



ML Cookbook

You need:

- A way of turning data -> training data
- A dataset
- **A model**
- Some infrastructure
- Evaluation metrics

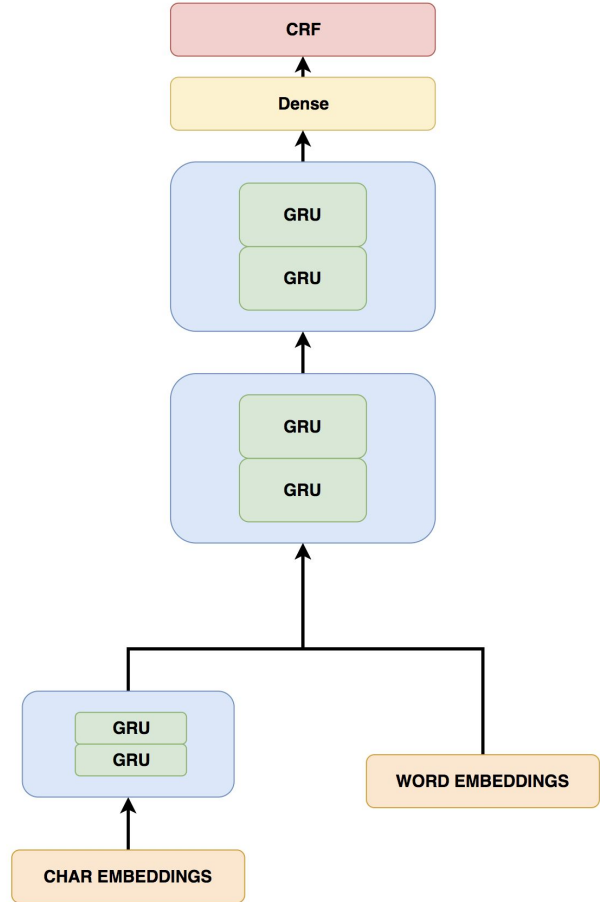


ML Model

Zhilin Yang & Salakhutdinov's Sequence Tagging Model:

- State-of-the-art at project inception
- BiGRU
- Learned character embeddings
- CRF decoding

What the heck is this thing?



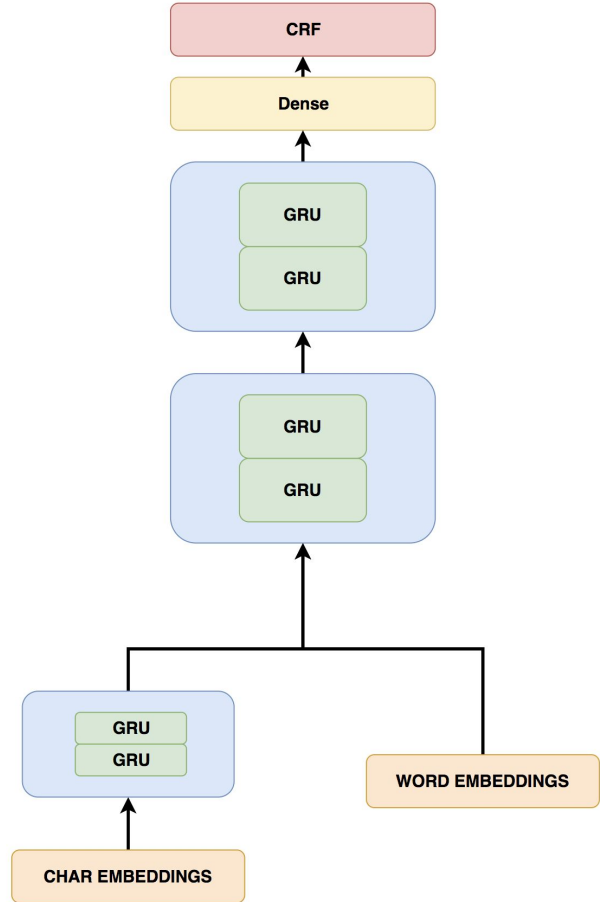
ML Model

Embeddings:

A set of k words can be embedded as a k -dimensional discrete space

But this makes it hard for the model to learn word associations (and results in huge model sizes)

We can project this k -dimensional space into a nicer low-dimensional vector space



ML Model

Embeddings:

We use skip-gram embeddings

Informally, embeddings are nice because they support addition and neural networks are good at doing addition

E.g. Paris - France + Poland = Warsaw

Paris - France is the idea of “Capital city”-ness

Distributed Representations of Words and Phrases and their Compositionality (Mikolov et al.)



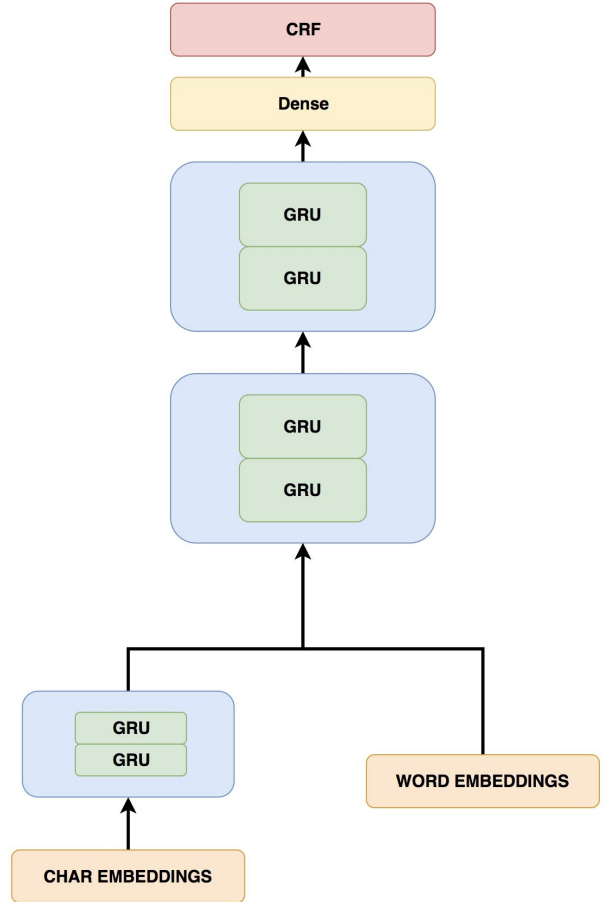
ML Model

Char Embeddings:

Learn an embedding for characters

This lets us get some information for words not in our vocabulary

Some new deep-learning models are entirely character-based



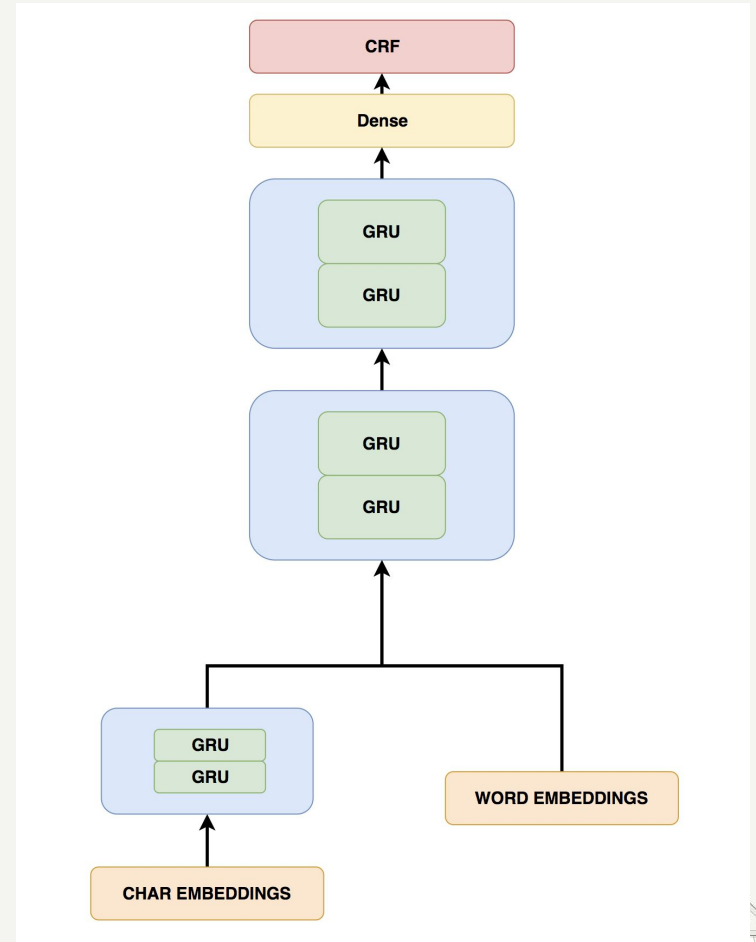
ML Model

GRU:

Inspiration: As you read a sentence, you store some information about what you have read so far

You use this information to infer the meaning of new words

After each word ask: "What is this sentence about?"



ML Model

GRU:

As the GRU reads, it gets its previous memory h_{t-1} and the current word x_t

It outputs its memory after each new word

We break our analogy by having the GRU read in both directions -- hence bidirectional

GRU is stacked two layers deep

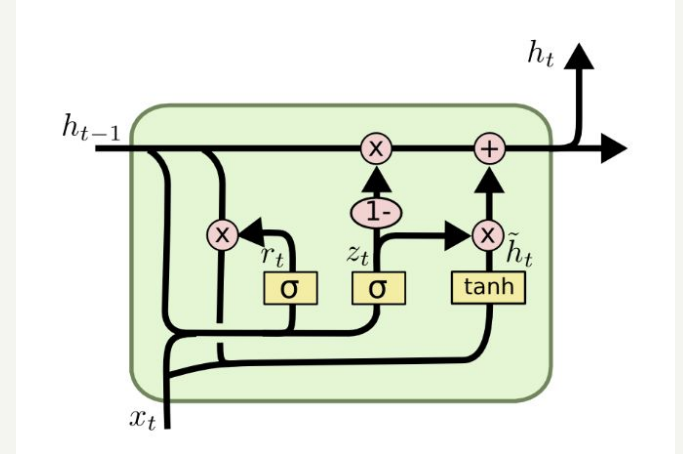


Image Taken from : <http://colah.github.io/>



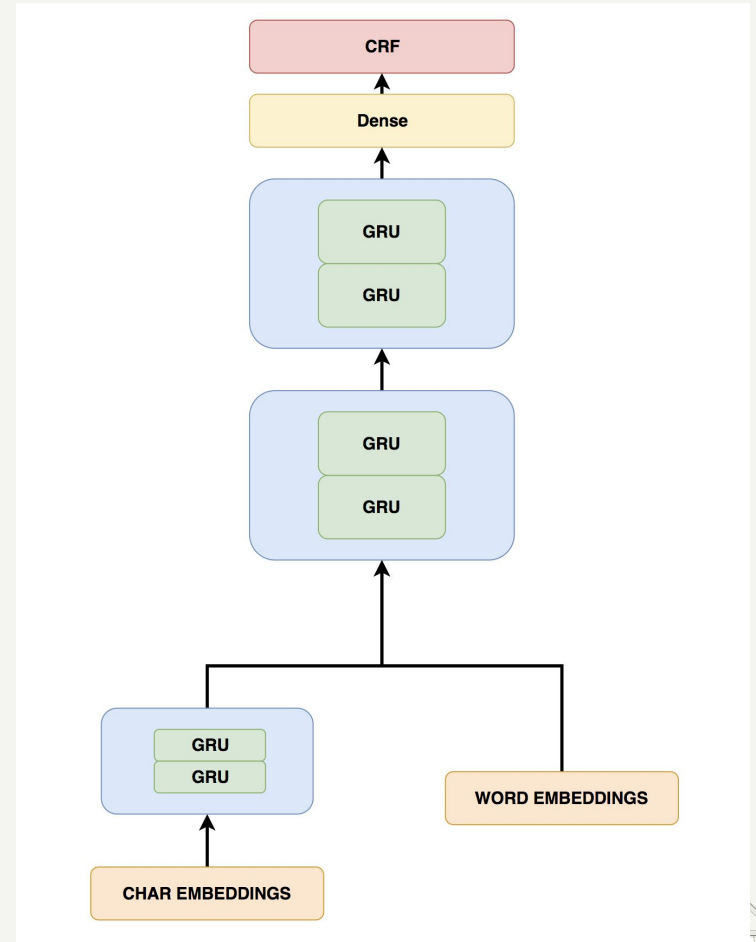
ML Model

CRF:

The dense layer gives us the probability that a word is a specific BIOES tag.

But taking the max probability isn't always correct

Also we might generate an invalid BIOES sequence (e.g. **BO**)



ML Model

CRF:

Model dependencies between labels

Given outputted labels $x_1 x_2 \dots x_{t-1}$, and predicted probabilities, find

$$X_t = \operatorname{argmax} p(L | x_1 x_2 \dots x_{t-1}) \text{ for all labels } L$$

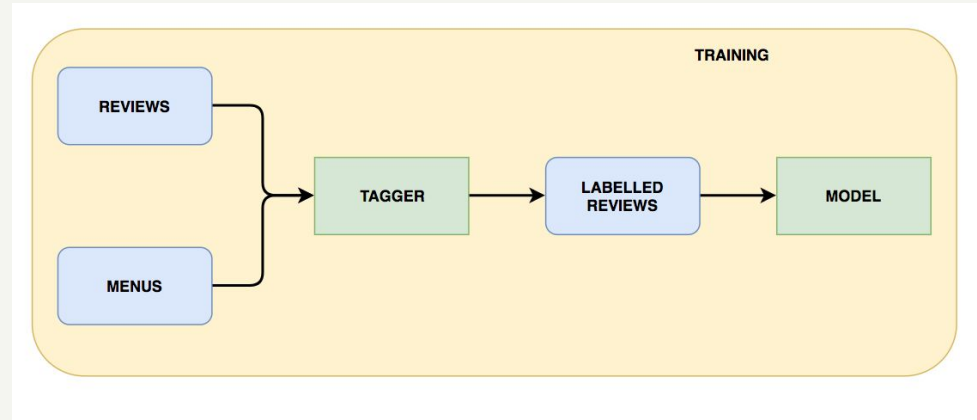
Do this by finding the probability that two labels are in sequence

$$\text{E.g. } p(\mathbf{E} | \mathbf{B}) = 0.7, p(\mathbf{O} | \mathbf{B}) = 0$$

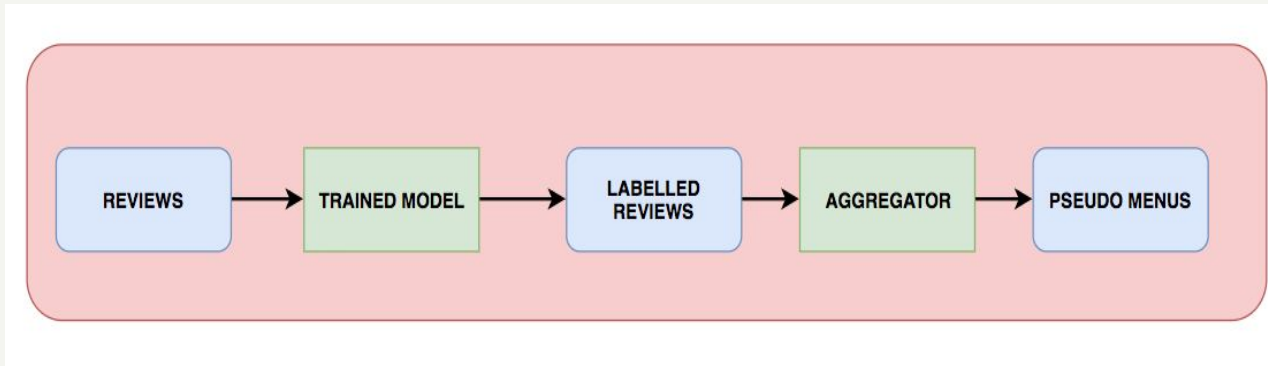


ML Model

We've described our model:



Now to create pseudomenus we aggregate our tags:



Aggregation Overview

Converts labelled reviews to Pseudomenus

We want to merge similar items:

- Fried Chicken Sandwich
- Crispy Fried Chicken Sandwich

We want to filter false positives



The Long Wait

★ 501 Reviews 📷 3 Photos

We use fuzzy matching to merge, frequency to filter



Aggregation Example

- **Labelled Reviews**

- The **Chicken Burrito** was good

- **Pseudo Menu**

- **Chicken Burrito** : 1

- **Not In Pseudo Menu**



Aggregation Example

- **Labelled Reviews**
 - The **Chicken Burrito** was good
 - The **Chickn Burrito** was good
- **Pseudo Menu**
 - **Chicken Burrito** : 2
- **Not In Pseudo Menu**
 - **Chickn Burrito** : Typo Duplicate



Aggregation Example

- **Labelled Reviews**

- The **Chicken Burrito** was good
- The **Chickn Burrito** was good
- The **Veggie Burrito** was good

- **Pseudo Menu**

- **Chicken Burrito** : 2
- **Veggie Burrito** : 1

- **Not In Pseudo Menu**

- **Chickn Burrito** : Typo Duplicate



Aggregation Example

- **Labelled Reviews**

- The **Chicken Burrito** was good
- The **Chickn Burrito** was good
- The **Veggie Burrito** was good
- The **Spicy Veggie Burrito** was good

- **Pseudo Menu**

- **Chicken Burrito** : 2
- **Veggie Burrito** : 2

- **Not In Pseudo Menu**

- **Chickn Burrito** : Typo Duplicate
- **Spicy Veggie Burrito** : Duplicate



Aggregation Example

- **Labelled Reviews**

- The **Chicken Burrito** was good
- The **Chickn Burrito** was good
- The **Veggie Burrito** was good
- The **Spicy Veggie Burrito** was good
- This restaurant was **good**

- **Pseudo Menu**

- **Chicken Burrito** : 2
- **Veggie Burrito** : 2

- **Not In Pseudo Menu**

- **Chickn Burrito** : Typo Duplicate
- **Spicy Veggie Burrito** : Duplicate
- **Good** : Low Confidence



Aggregation Example

After aggregation we'll get something that looks like this:

A pseudomenu!

```
[
  {
    "confidence": 0.6788991093635559,
    "tagged_count": 126,
    "name": "chocolate souffle"
  },
  {
    "confidence": 0.6851851940155029,
    "tagged_count": 78,
    "name": "roast maine lobster"
  },
  {
    "confidence": 0.6407766938209534,
    "tagged_count": 65,
    "name": "seared foie gras"
  },
  {
    "confidence": 0.8524590134620667,
    "tagged_count": 56,
    "name": "chocolate mousse"
  },
  {
    "confidence": 0.5925925970077515,
    "tagged_count": 52,
    "name": "glazed oysters"
  }
]
```



ML Cookbook

You need:

- A way of turning data -> training data
- A dataset
- A model
- **Some infrastructure**
- Evaluation metrics



ML Infrastructure

- How to get the dataset
- How to generate the training data
- How to train the model
- How to deploy the model



mrjob is Yelp's map-reduce framework

The main enemy of ML projects is time

One of the benefits of working at a large company is that these problems have been solved before :)



ML Cookbook

You need:

- A way of turning data -> training data
- A dataset
- A model
- Some infrastructure
- **Evaluation metrics**



Model Evaluation

Model metrics (per sentence):

- Per-class precision, recall, F-Score

Aggregation metrics (per pseudomenu)

- Precision, Recall, F-Score

But we're not confident that our training data is good, and we can't directly evaluate the user experience

Idea: A/B testing



Crowdsourcing Evaluation

Phrase:

chicken fingers

Context:

The real highlights were the rice - loaded with fresh vegetables, the chicken wings - crispy and meaty, the crab rangoon, **chicken fingers** - lightly battered and the beef teryaki - amazing. Traditional restaurant atmosphere that reminds me of my childhood. I was disappointed that it was so...

The highlighted phrase is (required)

- a food or drink
- not a food or drink

🔍 Make sure to look at the context. Some restaurants give their food weird names!

The highlighted phrase is (required)

- a main dish
- not a main dish

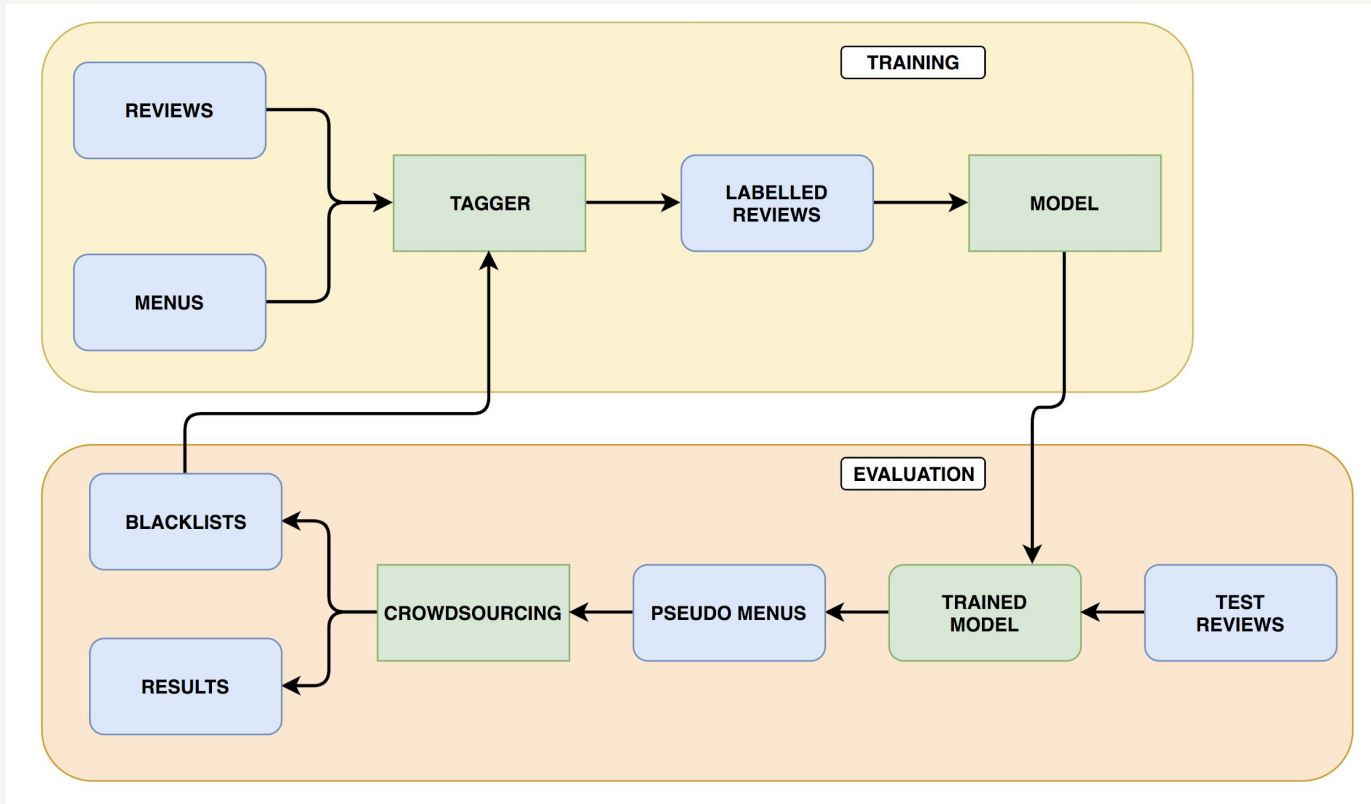
🔍 A "main dish" is something you could order at a restaurant as an entree or main dish. Appetizers, sides, desserts, ingredients, and drinks are not "main dishes."

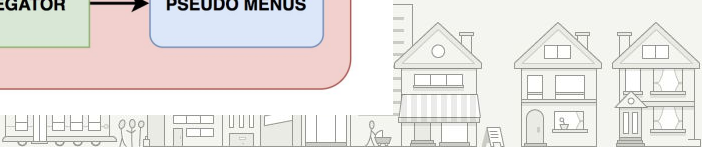
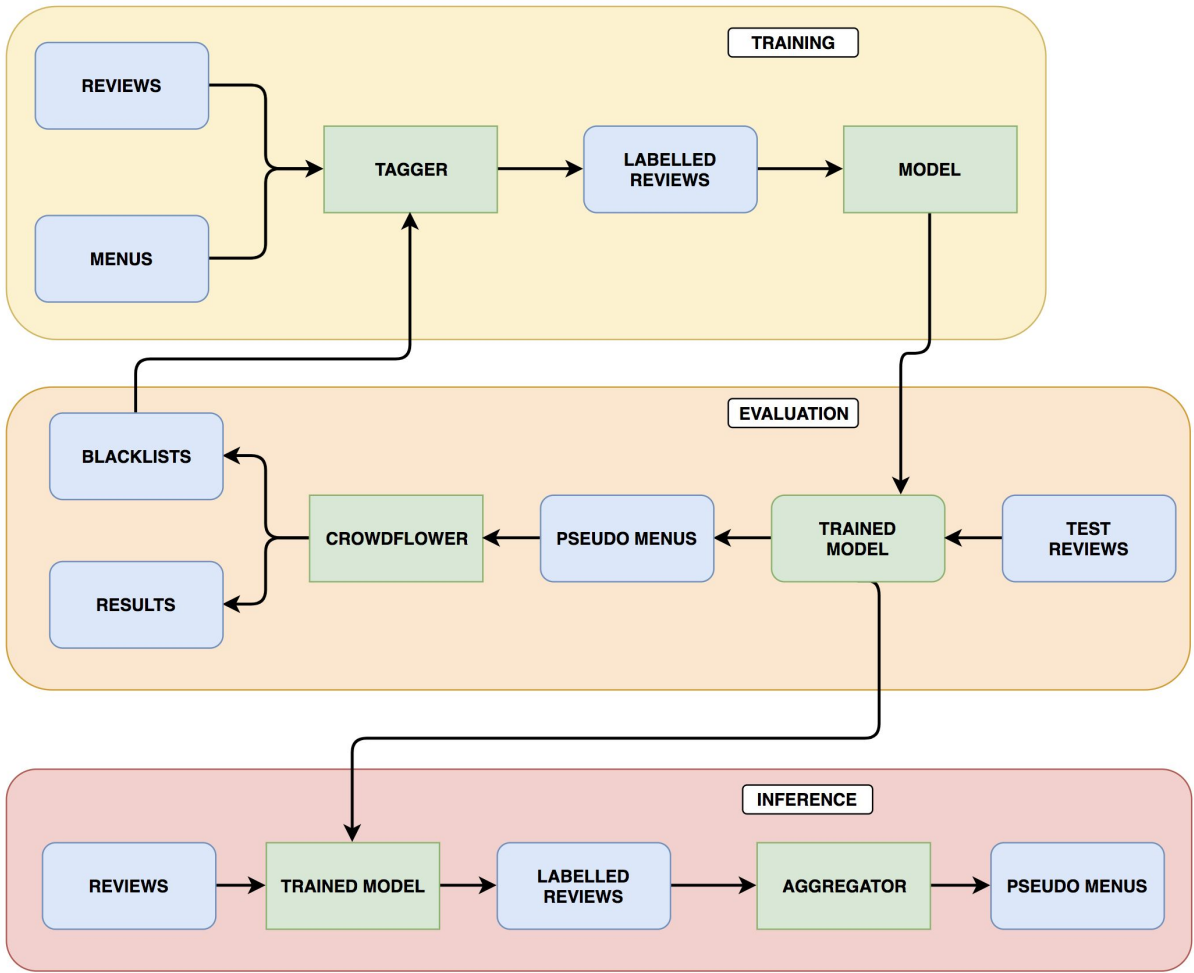
Check all duplicates (required)

- deviled eggs
- brussel sprout chips
- pork chop
- avocado toast
- marlowe burger
- crispy lemon chips
- demons on horseback
- soft poached egg
- smoked deviled eggs
- smoked salmon benedict
- NO DUPLICATES



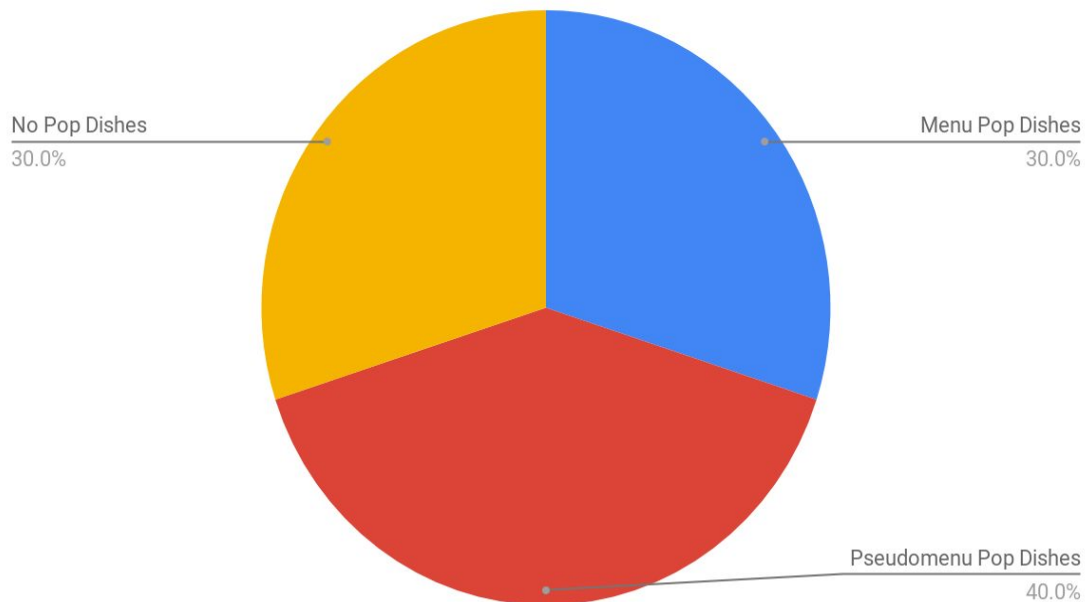
Human In the Loop Evaluation



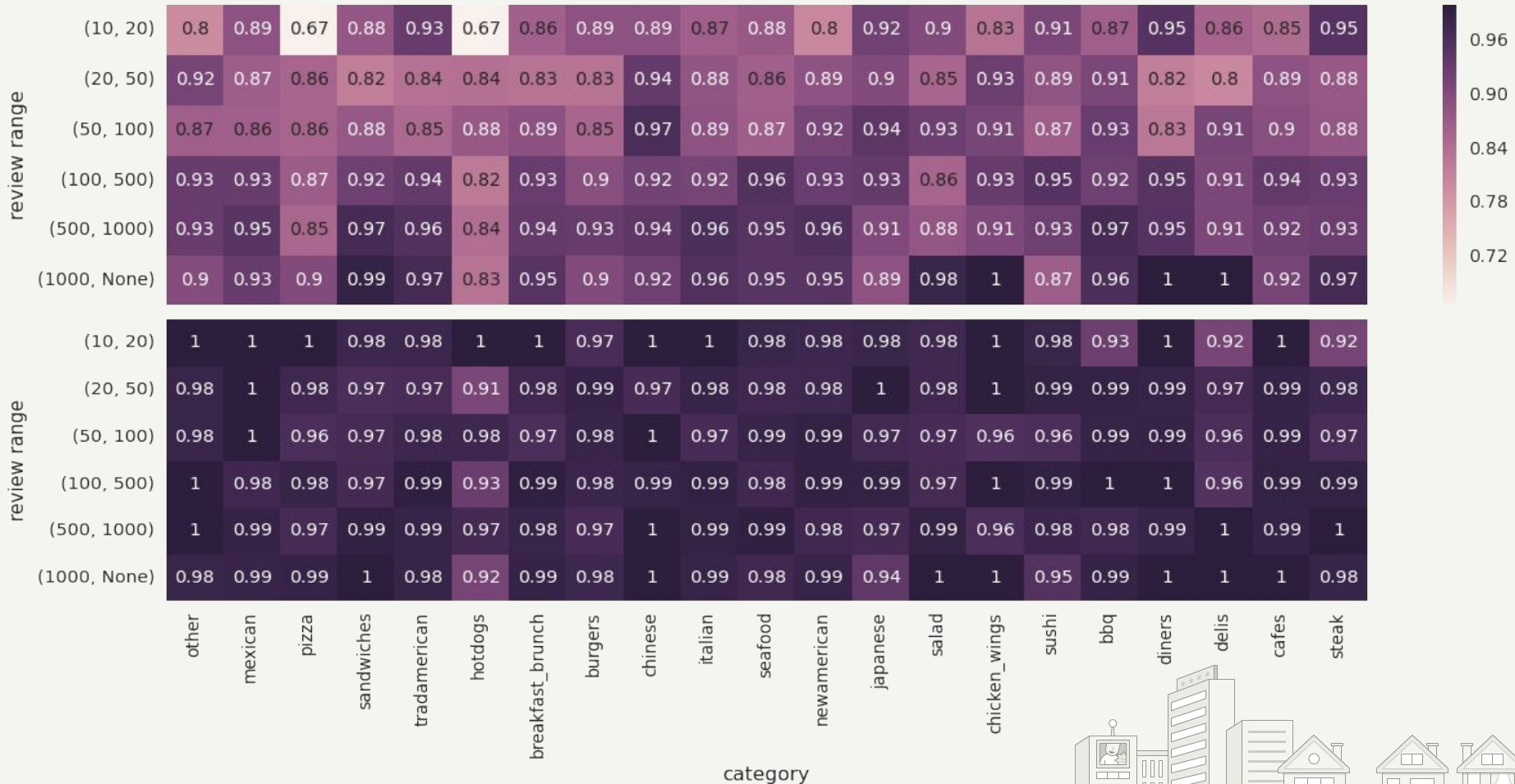


Coverage Evaluation

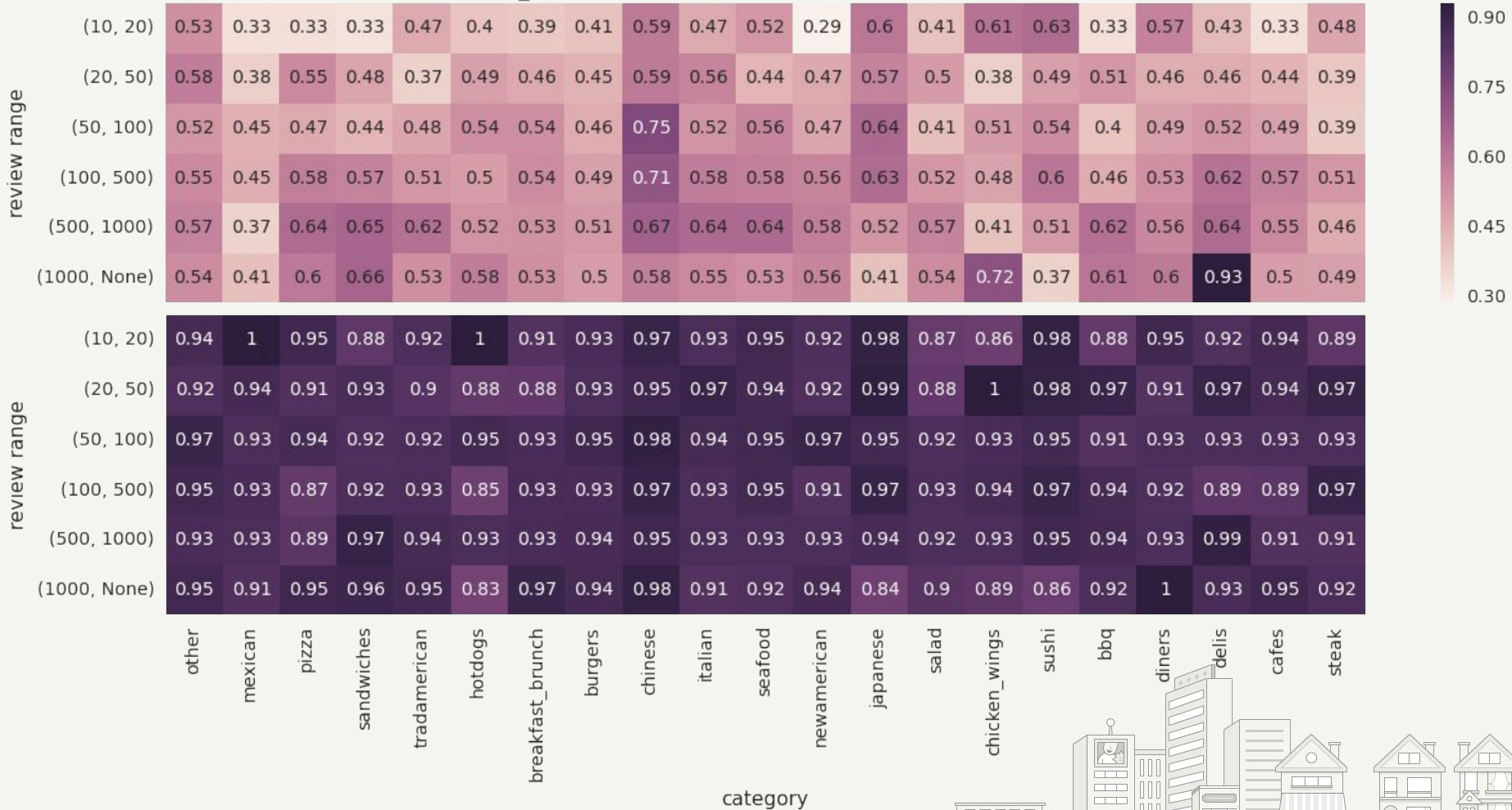
Popular Dishes Coverage



food accuracy in (category, review range) stratum

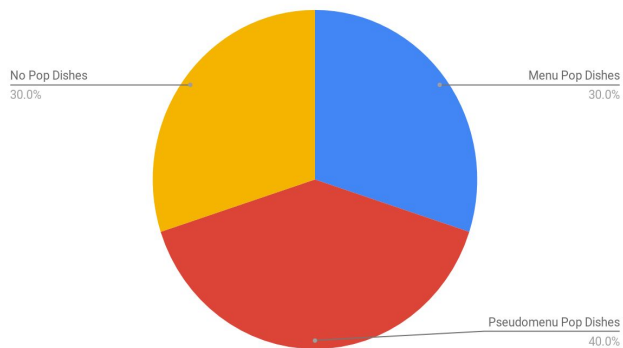


main_dish accuracy in (category, review range) stratum



Summary

Popular Dishes Coverage



Popular Dishes

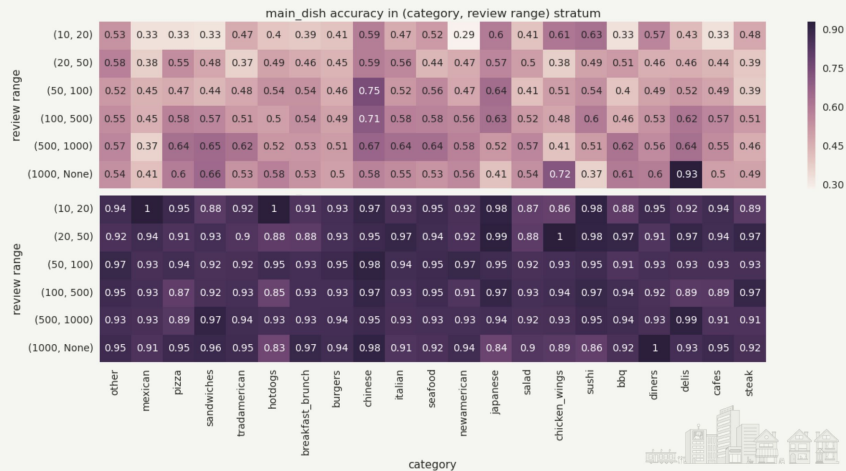


Teriyaki Chicken Savo...
16 Reviews • 6 Photos



Strawberry Banana
11 Reviews • 18 Photos

No-Menu Biz



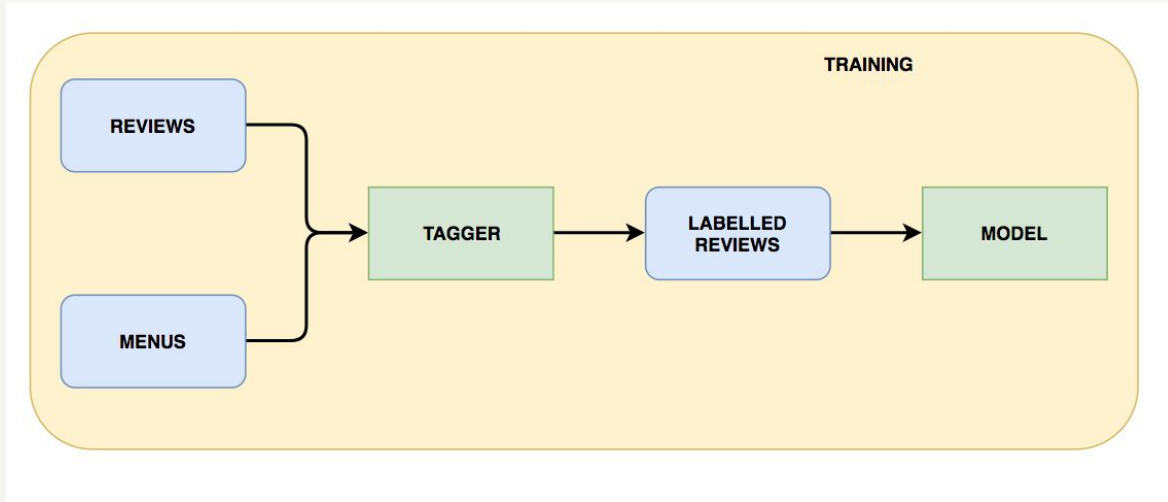
Improving Pseudomenus

- Improving the data
 - Data filtering
 - Data processing
- Improving the model
- Evaluating



Data Improvements

Earlier we saw:



The tagger's correctness directly affects our model
Our choice of Menus and Reviews is also important



Data Improvements

People don't always write what's exactly in the menu

A lot of work went into improving the robustness of the tagger

Fuzzy matching:

`The chicken burrito`	==	`chicken burrito`
`6 pieces chicken wings`	==	`chicken wings`
`flambé`	==	`flambe`
`chickn brurito`	==	`chicken burrito`
`n5. Broccoli Beef`	==	`broccoli beef`
`Chicken Curry ***`	==	`chicken curry`
`Veggie Burrito (vegan)`	==	`veggie burrito`



Data Improvements

Selecting Menus:

- **Idea:** Menus from different providers have quirks that cause them to be less suitable
- This causes food words to not be tagged in the training data, confusing the model

E.g. the menu item 'n5. Broccoli beef' is unlikely to appear in review text

Datasets:

1. All menus
2. Owner-verified only
3. Grubhub only



Data Improvements

This didn't work :(

Selecting Menus:

- Fuzzy-matching/regex was sufficient to deal with provider-specific “quirks”
- The decrease in train data size reduced performance
- **Lesson:** human prior knowledge can be used to clean training data
- Manually look through the training data to find patterns for Regex, etc.
- It's easy to find stray punctuation, bad prefixes, etc. and clean them



Data Improvements

Selecting Reviews:

Generally more data -> better model

But in this case out-of-date reviews sent to the tagger “poison” the training data

1. All reviews
2. Reviews from past 6 months
3. Reviews fresher than the menu
4. Intersection of #2 and #3

Best dataset: Past 6 months of reviews



Improving Pseudomenus

- Improving the data
 - Data filtering
 - Data processing
- Improving the model
- Evaluating



Transformer Networks

Introduced in *Attention Is All You Need* (Vaswani et al.)

Makes use of the attention mechanism first introduced in *Neural Machine Translation by Jointly Learning to Align and Translate* (Bahdanau, Cho and Bengio)

Had worse performance than our BiGRU until OpenAI released a paper on finetuning transformers



Transformer Networks

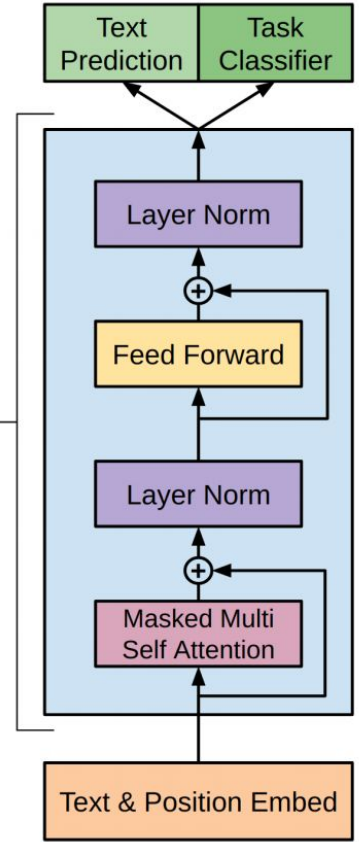
Finetuning:

Brings ideas of transfer learning to transformers:

If one person trains a transformer to perform a generic language task, everyone can go finetune it on domain-specific tasks

Radford et al. open-sourced their generic transformer for everyone (including us) to use!

12x



*Improving Language Understanding
with Unsupervised Learning*
(Radford et al.)



Transformer Networks

Main idea: Attention

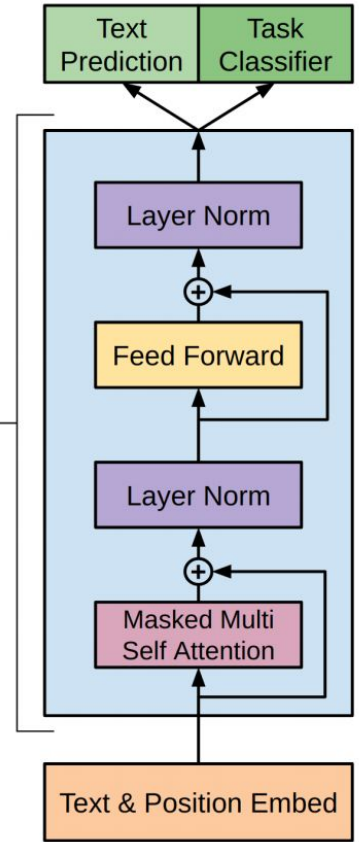
Given a sentence, attention asks “Which are the important words in the sentence?”

A weighted sum of those word embeddings acts like the “idea” of the sentence

But this seems like a difficult question to ask...

How do transformers deal with that?

12x



*Improving Language Understanding
with Unsupervised Learning*
(Radford et al.)



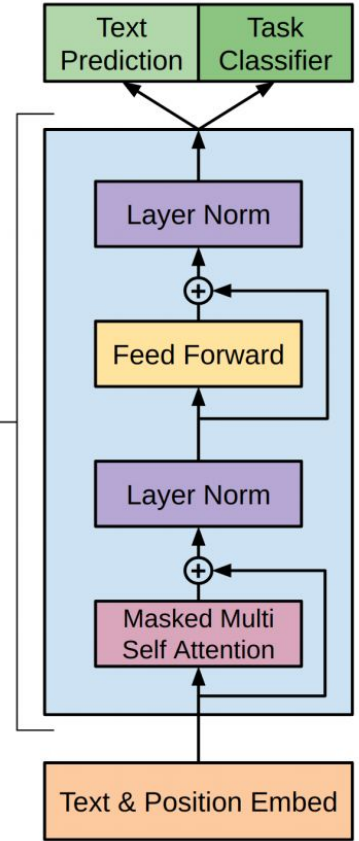
Transformer Networks

Masked Attention:

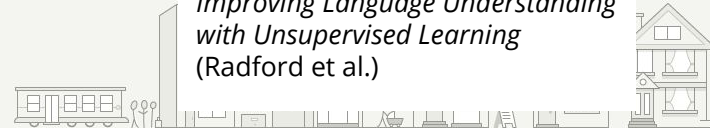
Have the model compute attention for every prefix of the sentence

This makes the attention simulate reading order

12x



*Improving Language Understanding
with Unsupervised Learning*
(Radford et al.)



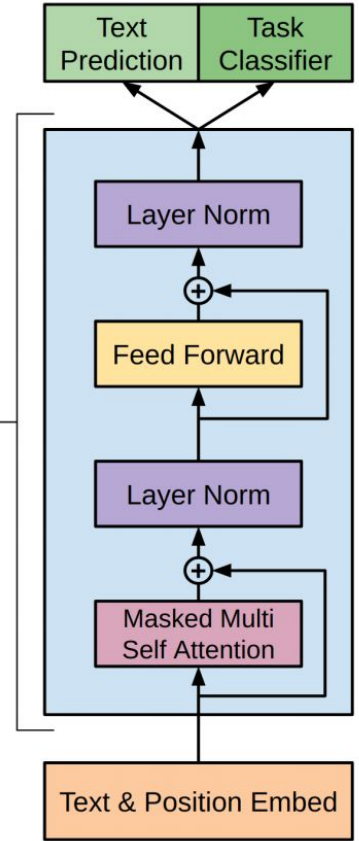
Transformer Networks

Multi-head Attention:

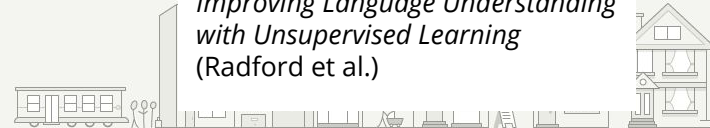
Instead of asking one question, ask several

This lets us split up the task of understanding sentence structure into several simpler tasks

12x



*Improving Language Understanding
with Unsupervised Learning*
(Radford et al.)



Transformer Networks

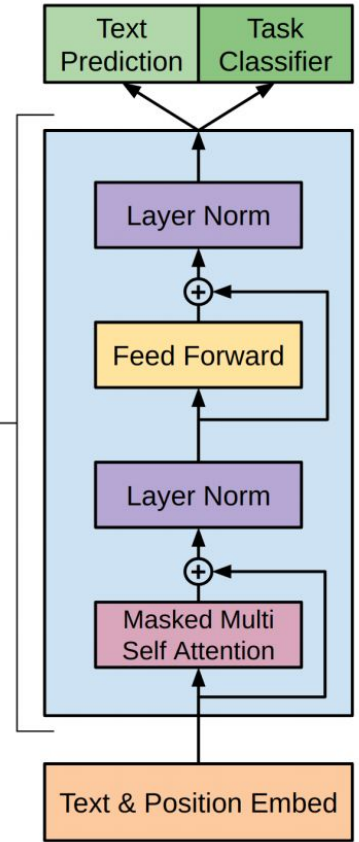
Self-attention:

Instead of computing a weight for each word, compute weights between pairs of vectors

Self-attention also has the sentence vector encode the question that should be asked

This is insanely cool! A layer tells the layer above it what kind of questions to ask about its structure

12x



*Improving Language Understanding
with Unsupervised Learning*
(Radford et al.)



Transformer Networks

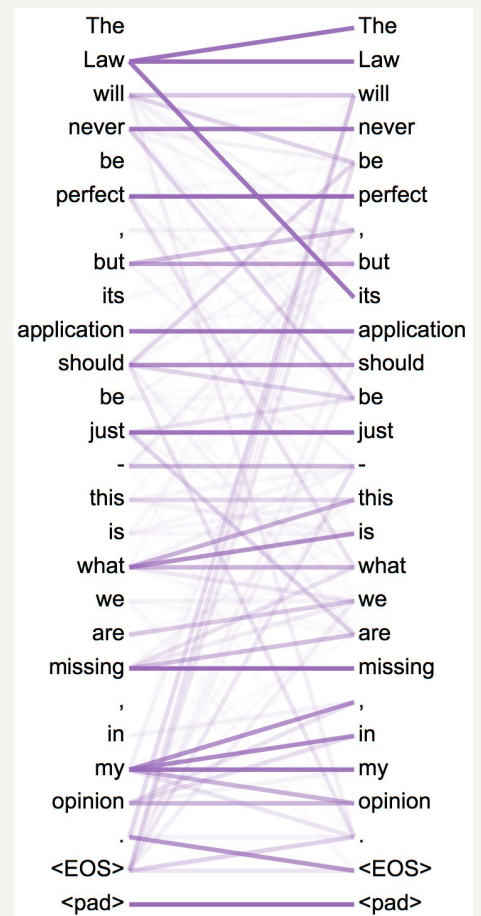
Attention Head Example:

We can visualize the action of attention heads

This head of the model learned performs anaphora resolution.

Anaphoras are when we use words to avoid repeating phrases.

- **Lucy** went to the movies. **She** had fun.
- **The Law** will never be perfect but **its** application should be just.



*Improving Language Understanding
by Generative Pre-Training*
(Radford et al.)



Improving Pseudomenus

- Improving the data
 - Data filtering
 - Data processing
- Improving the model
- Evaluating



Evaluating Results

Another CrowdFlower job

Phrase:

chicken fingers

Context:

The real highlights were the rice - loaded with fresh vegetables, the chicken wings - crispy and meaty, the crab rangoon, **chicken fingers** - lightly battered and the beef teryaki - amazing. Traditional restaurant atmosphere that reminds me of my childhood. I was disappointed that it was so...

The highlighted phrase is (required)

- a food or drink
- not a food or drink

🔔 Make sure to look at the context. Some restaurants give their food weird names!

The highlighted phrase is (required)

- a main dish
- not a main dish

🔔 A "main dish" is something you could order at a restaurant as an entree or main dish. Appetizers, sides, desserts, ingredients, and drinks are not "main dishes."

Non-main dish frequency decreased by 25%
Number of popular dishes increased by 16%



Miscellaneous Improvements

Gold Sentence	Predicted Sentence
This is the first time in a long time but I actually received real smoked bacon , you can taste the smoke flavor throughout .	This is the first time in a long time but I actually received real smoked bacon , you can taste the smoke flavor throughout .
The servers really know the menu and are able to help you choose the perfect dish .	The servers really know the menu and are able to help you choose the perfect dish .
Great ! Pad Thai is perfectly made with real chicken .	Great ! Pad Thai is perfectly made with real chicken .
The Hawaiian Roll was out of this world .	The Hawaiian Roll was out of this world .
I 'm not a big bleu cheese fan but when it 's creamed in spinach , it 's awesome .	I 'm not a big bleu cheese fan but when it 's creamed in spinach , it 's awesome .
I have never had a better burger ANYWHERE !	I have never had a better burger ANYWHERE !
If you 're a fan of really crispy and tasty catfish , this is the place for you ! !	If you 're a fan of really crispy and tasty catfish , this is the place for you ! !
I say come here once for drinks but back again for duck when you are in the mood .	I say come here once for drinks but back again for duck when you are in the mood .
My mom and I shared the turkey provolone and bacon sandwich and my dad and my husband had the roast beef sandwiches .	My mom and I shared the turkey provolone and bacon sandwich and my dad and my husband had the roast beef sandwiches .
Good place to eat with family and friends I love their food my son like the chicken lollipops and I like chilli chicken	Good place to eat with family and friends I love their food my son like the chicken lollipops and I like chilli chicken
Hubby enjoyed the clam chowder and I had the baked cod .	Hubby enjoyed the clam chowder and I had the baked cod .
Pesto and Bolognese were good .	Pesto and Bolognese were good .
Some of my teammates ordered the Shawarma Chicken Sandwich , the Falafel Sandwich , the Shish Taouk Platter , Grilled Lamb Chops platter along with Lentil Soup and Rice and Lentils .	Some of my teammates ordered the Shawarma Chicken Sandwich , the Falafel Sandwich , the Shish Taouk Platter , Grilled Lamb Chops platter along with Lentil Soup and Rice and Lentils .
My wife and I have been coming to Lulu for the past 4 years .	My wife and I have been coming to Lulu for the past 4 years .
I ordered the chicken shawarma plate with a side of falafel .	I ordered the chicken shawarma plate with a side of falafel .
If you have an issue with sharing space , this is not the place for you .	If you have an issue with sharing space , this is not the place for you .
We shared orders of Ugali Sticks (\$ 7) , essentially fried polenta sticks with coconut and Applewood - smoked clover honey ; 24-Hour Jerk Wings (\$ 12) ; and Smoke - Braised Oxtail (\$ 24) .	We shared orders of Ugali Sticks (\$ 7) , essentially fried polenta sticks with coconut and Applewood - smoked clover honey ; 24-Hour Jerk Wings (\$ 12) ; and Smoke - Braised Oxtail (\$ 24) .
The latter orders of pork belly were much better so your meat may vary .	The latter orders of pork belly were much better so your meat may vary .
It melted in my mouth !	It melted in my mouth !
For my main course , I went with the fish tacos on the recommendation of my server .	For my main course , I went with the fish tacos on the recommendation of my server .
My favorite new dessert is the cocoa - nut dream .	My favorite new dessert is the cocoa - nut dream .
We shared the chicken taco , trinity taco , the fido taco(which for being vegetarian was arguably # 1 or # 2) We also had the Guadalupe taco and shared the Andy Migas .	We shared the chicken taco , trinity taco , the frio taco(which for being vegetarian was arguably # 1 or # 2) We also had the Guadalupe taco and shared the Andy Migas .



Miscellaneous Improvements

- Debug view (previous slide)
- Improved evaluate metric (fuzzy F-score)
 - Subset matching
 - **Gold:** I had the chicken **burrito**
 - **Predicted:** I had the **chicken burrito**
 - Common food matching
 - **Gold:** I had the nachos
 - **Predicted:** I had the **nachos**



Summary

ML isn't always the way to go

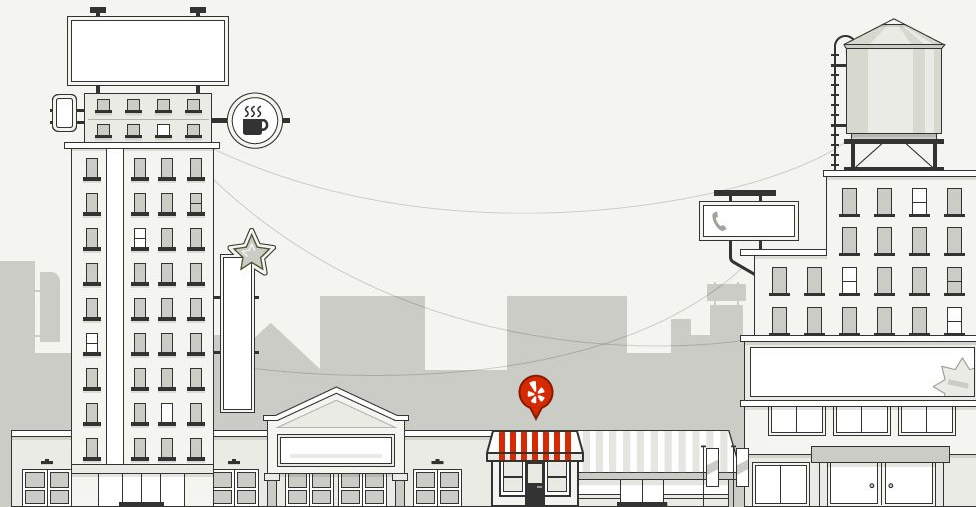
Your main enemy in ML projects is time

Exploring the output of your model/scripts helps you ideate

Be creative! Come up with metrics that model your problem



Questions





We're Hiring!

www.yelp.com/careers/



fb.com/YelpEngineers



[@YelpEngineering](https://twitter.com/YelpEngineering)



engineeringblog.yelp.com



github.com/yelp