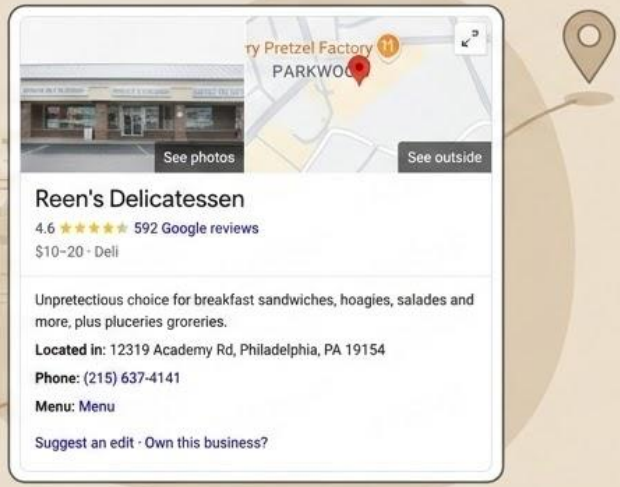
Project B: Places Attribute Conflation

Presented by: Stanley Shen, Jeffrey Liu

Project Overview

What is the problem?

When one real-world place has conflicting attributes (name, address, etc.) across multiple sources, which one is correct?



Significance to OMF and community

High-quality, unified data is foundational for OMF

Conflation ensures a single, trustworthy record for every place

Reducing user confusion and maximizing data utility for the wider community (e.g., navigation, business analysis)

OKRS

OKR1:

Objective 1: Build a high-quality labeled dataset for attribute conflation

Key Results:

- Collect and preprocess ≥2000 pre-matched place records from ≥2 distinct data sources.
- Label ≥90% (≥1800) of collected records with verified ground-truth attributes using Yelp Academic Dataset.
- Achieve ≥95% data consistency across the final labeled dataset via automated field validation.
- Ensure ≤2% missing or invalid attribute values (≤40 records) after final cleaning.

OKRS

OKR2:

Objective 2: Develop and compare algorithms for optimal attribute selection

Key Results:

- Implement 2 algorithms: one rule-based and one machine learning model.
- Achieve ≥80% validation accuracy on attribute prediction using a 70/15/15 train/validation/test split.
- Limit model inference time to ≤200 ms per record on 2000 records.

OKRS

OKR3:

Objective 3: Evaluate performance and recommend a scalable solution

Key Results:

- Evaluate model performance using precision, recall, and F1-score on a test set of 2000 records.
- Demonstrate ≥5% relative improvement in F1-score of the best-performing model compared to the baseline rule-based approach.
- Provide ≥3 recommendations for scaling the solution to datasets exceeding 1 million records.

OKR Journey

- Suggested to start with 2k-3k data samples instead of 10k
- Consider a) model performance and (b) model trade-offs between a linear /rule based system and a non-linear ml-based model. Are there shortcomings that both model variants share?

We found that:

- Used non-linear ML based since there are a lot of attributes, linear ml based would be like rule based
- Hybrid approach: use RandomForest for entity matching, rules for conflation

Approach

- **Phase 1 – Data Preparation**
  - Collect place data from multiple sources (Yelp, OMF, OpenStreetMap, etc.)
  - Clean and normalize data: standardize text, remove duplicates
- **Phase 2 – Entity Matching / Place ID Assignment**
  - Generate candidate record pairs using city, ZIP, or proximity
  - Match records via fuzzy name/address, category, and optional lat/lon
  - Assign unified place_id and evaluate matches against known data

place_id	source	name	address	lat	lon
P0001	Yelp	Starbucks Coffee	123 Main St	…	…
P0001	OMF	Starbucks Coffee	123 Main Street	…	…
P0001	Overpass	Starbucks	125 Main St	…	…

Ruled based

The Methodology: We implemented a three-step Python pipeline:

1. **Normalize:** Cleaning text and formatting (e.g., stripping non-digits from phone numbers) to ensure fair comparisons.
2. **Rule Engine:** Applying logical heuristics, such as selecting the longest valid address or using majority voting for phone numbers.
3. **Validate:** Scoring the output against a human-verified dataset.

place_id	source	name	address	lat	lon
P0001	Yelp	Starbucks Coffee	123 Main St	…	…
P0001	OMF	Starbucks Coffee	123 Main Street	…	…
P0001	Overpass	Starbucks	125 Main St	…	…

Hybrid

- Use ML to entity match
  - Pure Rules fail at matching (misses fuzzy duplicates).
  - Name, address, distance similarity between pairs of the same POI from different data sets (ex Yelp, OMF), matching is non-linear so use RandomForest
- Use rule based to conflate
  - Pure ML fails at selection (hallucinates fake names).

Machine Learning

The Methodology: We implemented a supervised learning pipeline:

1. **Feature Engineering:** Converting raw text into numerical features (e.g., token similarity and data presence) to enable algorithmic processing.
2. **Model Training:** Training Random Forest and Logistic Regression classifiers to predict the most reliable source based on patterns in our ground truth.
3. **Validate:** Scoring the model predictions against our human-verified dataset to measure improvement over the rule-based baseline.

Metrics/Demo

	Rule Based	Hybrid	Machine Learning
Entity Matching Accuracy	X	*93.78*	X
Conflation Accuracy	80.10	*90.80*	84.6
		Overall: 85.15	

https://github.com/project-terraforma/stanley-jeffrey-attributesConflation/blob/main/Results.md

Reflection/Takeaway

Algorithmic Trade-offs: Rule-based logic offers explainability and speed for general cases, while Machine Learning provides superior adaptability for complex conflicts at the cost of training data requirements.

The Data Foundation: Both approaches demonstrated that high-quality Entity Resolution is impossible without a "Human-in-the-Loop" to establish the Ground Truth for validation and training.

Resources

Repository:
https://github.com/project-terraforma/stanley-jeffrey-attributesConflation

Additional Resources:

https://docs.overturemaps.org/

https://overpass-turbo.eu/

https://business.yelp.com/data/resources/open-dataset/

Thank you