Comparing learning apporaches for entity resolution in public transport schedule data

Research proposal

Mitchell McMillan

12 March 2026

Motivation and question

- **Motivation:** City-scale public transport schedule datasets provide the universe of services - potentially rich insights into public transport access across space and time with high granularity (every origin location, every destination, every second).
- Quality of schedule data is aimed at navigation use cases: missing trips are much more problematic than duplicate trips.
  In London context, cost-constrained public authority + small, technically unsophisticated private providers $\implies$ lots of (potentially *fuzzy*!) duplicates $\implies$ entity resolution (ER) problem.
- **Question:** Can active learning minimise labelling effort while maintaining accuracy in identifying duplicate trips in public transport datasets?
- **Economic relevance:** Duplicate trip records distort measured service frequency, and thus capacity, travel times and accessibility.

Literature and contribution

- Active learning can reduce the number of labels needed relative to passive learning - relevant when labels are costly (Li et al. 2025).
- Active learning has been applied to large scale ER problems (Binette and Steorts 2022; Qian, Popa, and Sen 2017)
- Diverge from standard ER by extending label to:
  - keep both
  - keep A
  - keep B
- Aim to compare (1) passive, (2) cold-start active, and (3) warm-start active learning for public transport trip deduplication in GTFS datasets.

Evaluation

- **Methods**
  1. Unsupervised baseline: Duplicate detection (mixture model) + retention decision (heuristic trip quality score $\rightarrow$ k-means on difference in scores)
  2. Passive supervised learning: training up-front on training sample
  3. Cold-start active learning: adaptively query labels from an initially unlabelled training pool
  4. Warm-start active learning: begin with a small labelled seed sample, then query adaptively
- **Training:** Supervised and active-learning methods use training sample of labelled set
- **Evaluation:** All methods are evaluated on test sample of labelled set
- **Outcomes:**
  - action accuracy on the test sample
  - number of labels used in training
  - active model loss: $L_t = \mathbf{1}\{\text{wrong action}\} + \lambda \cdot \mathbf{1}\{\text{training label used}\}$

Feedback?

References I


Binette, Olivier and Rebecca C. Steorts (2022). "(Almost) all of entity resolution". In: *Science Advances* 8.12, eabi8021. DOI: 10.1126/sciadv.abi8021. eprint: https://www.science.org/doi/pdf/10.1126/sciadv.abi8021. URL: https://www.science.org/doi/abs/10.1126/sciadv.abi8021.


Li, Dongyuan et al. (2025). "A Survey on Deep Active Learning: Recent Advances and New Frontiers". In: *IEEE Transactions on Neural Networks and Learning Systems* 36.4, pp. 5879–5899. DOI: 10.1109/TNNLS.2024.3396463.


Qian, Kun, Lucian Popa, and Prithviraj Sen (2017). "Active Learning for Large-Scale Entity Resolution". In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. CIKM '17. Singapore, Singapore: Association for Computing Machinery, pp. 1379–1388. ISBN: 9781450349185. DOI: 10.1145/3132847.3132949.

Appendix: current static pipeline

- **Stage 1: identify duplicate-like pairs**
  - Remove exact duplicates and generate candidate overtaking pairs.
  - Construct pair features from stop-by-stop timing differences, including exact-match share, gap size, stability of the gap, and sign changes across stops.
  - Fit an unsupervised Gaussian mixture model to these features.
  - Interpret the cluster with smaller average timing gap as the duplicate-like cluster.
- **Stage 2: choose which trip to retain**
  - For duplicate-like pairs, assign each trip a quality score based on timetable plausibility.
  - Prefer the trip with the better quality score.
  - Use k-means on quality margins to decide whether to auto-drop the weaker trip or flag the pair for manual review.