

Proceedings of Statistics Canada Symposium 2024: The Future of Official Statistics

Efficient Record Linkage for Large Datasets by Business Names

by Hanan Ather, Serge Godbout, and Dave MacNeil

Release date: September 8, 2025



Efficient Record Linkage for Large Datasets by Business Names

Hanan Ather, Serge Godbout, and Dave MacNeil¹

Abstract

Accurate and efficient record linkage is crucial for maintaining a comprehensive and current Statistical Business Register (SBR) at Statistics Canada. Linking external business lists to the SBR by name presents computational and methodological challenges, especially as data volumes grow. This paper describes a scalable methodology that employs blocking techniques to constrain the computational search space and integrates multiple similarity measures—from edit distances and n-gram overlaps to embedding-based methods using Sentence-BERT (SBERT)—to identify likely matches. By combining simple character-level comparisons with more advanced semantic embedding methods, the approach can adapt to various naming conventions and complexities. While it does not guarantee superior accuracy in all circumstances, it offers a pragmatic balance between computational feasibility and linkage quality.

Key Words: Data integration; Record linkage; Big data; Deep learning.

1. Introduction

1.1 Background and Context

The Statistical Business Register (SBR) at Statistics Canada supports a wide range of economic surveys and analytical programs. It serves as an authoritative frame of businesses operating within Canada, regularly updated to ensure coverage and representativeness. Maintaining the quality of the SBR involves integrating new data sources that detail changes in the business landscape, be they administrative files, external listings, or other reference datasets. A central challenge in this integration is ensuring that business entities from disparate sources can be accurately linked to their corresponding entries in the SBR. Although business names provide a common linkage key, variations in spelling, abbreviations, formatting, and typographical errors limit the effectiveness of exact string-matching approaches.

1.2 Need for Scalable Linkage Methods

As the volume and variety of source data grow, the computational demands associated with record linkage become more pronounced. Simple methods may suffice for small or relatively uniform datasets, but large-scale linking—potentially involving millions of records—requires more advanced methods. Without careful design, the number of pairwise comparisons can increase quadratically, making the process costly and time-consuming. Thus, a key objective is to develop a strategy that not only captures complex name variations but does so efficiently, ensuring that resource demands remain within practical limits.

1.3 Multi-Stage Linkage Approach

The methodology presented in this paper adopts a multi-stage framework designed to balance computational feasibility and the reliability of matched pairs. The first stage employs a blocking procedure to partition records into smaller subsets that share certain attributes, such as geographic identifiers. By limiting subsequent comparisons to these

¹**Hanan Ather**, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Canada, K1A 0T6 (hanan.ather@statcan.gc.ca); **Serge Godbout** (serge.godbout@statcan.gc.ca), Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Canada, K1A 0T6; **Dave MacNeil** (dave.macneil@statcan.gc.ca), Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Canada, K1A 0T6.

subsets, the number of candidate pairs is dramatically reduced. The second stage applies a suite of similarity measures to these candidate pairs. Character-level methods, such as edit distance and n-gram overlaps, quantify textual resemblance, while embedding-based approaches leverage Sentence-BERT (SBERT) to capture semantic relationships that transcend simple string matching. Finally, threshold-based decision rules determine which pairs qualify as matches, allowing analysts to adjust parameters to emphasize precision or recall depending on the requirements of their particular use case.

1.4 Objectives and Contributions

The goal of this work is to provide a methodology that supports the integration of external business data into the SBR on a scale that aligns with real-world data volumes. Rather than claiming to surpass all existing methods in accuracy, the paper focuses on practical trade-offs that yield meaningful efficiency gains. By combining blocking strategies with progressively more sophisticated similarity measures, it aims to facilitate the use of embedding-based techniques in settings where naive, exhaustive comparisons would be infeasible. This approach offers analysts a configurable linkage process that can accommodate varying data conditions, resource constraints, and quality expectations.

2. Background and related work

2.1 The role of the Statistical Business Register (SBR)

The Statistical Business Register (SBR) at Statistics Canada serves as an essential infrastructure for the production of economic statistics. Compiled from multiple data sources and regularly updated, it provides a comprehensive frame of businesses across the country. Many programs that rely on the SBR, including over two hundred economic surveys, depend on the ability to maintain an accurate and current representation of the business population. Ensuring this accuracy requires integrating new datasets, refining records over time, and reconciling information from disparate administrative files or external registries (Oyarzun and Wile, 2016).

2.2 Balancing complexity, scalability, and accuracy

Linking external records to the SBR by business name is challenging due to variations in representation and formatting. These range from simple spelling differences to more complex discrepancies, such as the use of abbreviations or partial corporate descriptors. Exact string matching rarely suffices, as legitimate counterparts may not share identical text. Conventional linking approaches, while successful in smaller or more homogenous datasets, can struggle to scale efficiently to large files, becoming computationally expensive and increasingly unwieldy as data volumes grow. When dealing with millions of potential matches, it becomes critical to adopt methods that can manage complexity, maintain reasonable levels of accuracy, and execute in a timely manner. The approach described here aims to offer a more computationally manageable solution to link large datasets through incremental methodological enhancements, acknowledging that it is neither flawless nor necessarily superior in all cases, but that it can improve efficiency and scalability within practical constraints.

3. Methodology

3.1 Overview of the framework

This innovative record linkage method attempts to find pairs (x, y) where $x \in X$ is from the SBR and $y \in Y$ is from an external dataset, that refer to the same underlying entity. Each pair is assigned a similarity score $s(x, y) \in [0, 1]$ and pairs exceeding a threshold t are considered matches. Directly comparing all pairs is often impractical at large scales, prompting the use of a blocking step that partitions the datasets into smaller subsets. Similarity measures ranging from simple edit distances to SBERT embeddings are then applied, allowing the most computationally expensive operations to be focused where they are needed.

3.2 Blocking for computational efficiency

Blocking reduces the search space before similarity measures are computed. A function B assigns records to blocks based on attributes such as geographic information. Only pairs (x, y) falling into the same block are compared, transforming a potentially quadratic process into a manageable sum of within-block operations. Although blocking may miss some distant matches, it significantly improves scalability and enables the subsequent use of more complex similarity methods.

3.3 String-based similarity measures

Character-level and substring-based metrics remain essential components of the linkage pipeline. They provide a baseline level of textual comparison at relatively low computational cost.

3.3.1 Edit distance (Levenshtein distance)

The Levenshtein edit distance $d_{LEV}(x, y)$ measures the minimum number of single-character edits needed to transform one string into another (Levenshtein, 1966). The similarity score is derived as:

$$s_{LEV}(x, y) = 1 - \frac{d_{LEV}(x, y)}{\max(|x|, |y|)}$$

This measure handles simple spelling variations and typographical errors but is sensitive to word order changes and larger structural differences.

3.3.2 N-gram

N-gram methods represent strings as sets of overlapping substrings of n . Set based methods are also used for numerically quantifying the similarity between two strings. The two strings are viewed as sets of tokens and the similarity of two strings x and y is a measure of their “overlap”. Let T_x be the set of tokens generated by string x , and T_y be the set of tokens generated by string y . We define the overlap of the two sets by $O(x, y) = |T_x \cap T_y|$. For example, let $x = \text{'STATCAN'}$ and $y = \text{'STATISTICS CANADA'}$; $T_x = \{\text{'ST', 'TA', 'AT', 'TC', 'CA', 'AN'}\}$ and $T_y = \{\text{'ST', 'TA', 'AT', 'TI', 'IS', 'ST', 'TI', 'IC', 'CS', 'S', 'C', 'CA', 'AN', 'NA', 'AD', 'DA'}\}$. Then,

$$O(x, y) = |T_x \cap T_y| = \{\text{'ST', 'TA', 'AT', 'CA', 'AN'}\}$$

And respectively.

$$S_{NGRAM}(x, y) = \max\left(0, \frac{O(x, y)}{\text{String Length}}\right)$$

3.4 Embedding-based similarity measures

In Natural Language Processing (NLP), embeddings are numerical representations of text that facilitate machine understanding. Text is encoded into a multi-dimensional vector space, where each dimension captures specific linguistic or contextual properties. In this vector space, the position of a word is determined by its numerical values across multiple dimensions. Words with similar meanings or contexts are mapped closer together, enabling the model to capture semantic relationships between them. The output of the embedding model corresponds to a specific location in this multi-dimensional space, representing the text's underlying semantic meaning. Embeddings are crucial for capturing semantic meaning, encompassing both the context and intent of the text. Transformer models, leveraging embeddings, are more capable of recognizing that these variations refer to the same entity, even when word order is altered, by capturing deeper semantic understanding.

3.4.1 Sentence-BERT (SBERT)

SBERT is a variant of the BERT (Bidirectional Encoder Representations from Transformers) model specifically optimized for generating semantically meaningful sentence embeddings (Reimers and Gurevych, 2019). Instead of relying on a masked language modeling objective alone, SBERT uses a siamese network architecture where two

transformer networks share weights and are trained on a sentence-pair classification or regression task. This training encourages similar sentences (in terms of meaning) to have embeddings that are close in vector space.

Given a business name x , we first preprocess it (e.g., uppercase conversion, whitespace normalization) and then feed it into the SBERT model. The SBERT encoder maps x into a vector $v(x) \in R^d$, where d is the embedding dimension (commonly in the hundreds). the embeddings $v(x)$ and $v(y)$ should be close in R^d -space, reflecting the similar meaning.

3.4.2 Cosine Similarity for Embeddings

To measure the similarity between two embedded names, $v(x)$ and $v(y)$, we use the cosine similarity. Cosine similarity is defined as

$$s_{cos}(x, y) = \frac{v(x) \cdot v(y)}{\|v(x)\| \|v(y)\|}$$

Table 3.4.2.1-1 demonstrates how different scoring methods handle word order when comparing strings, using "HANAN ATHER TRUCKING INC" as the canonical reference.

**Table 3.4.2.1-1
Comparison of Different Scores**

| BR Name | Levenshtein Score | 1-gram score | 2-gram score | 3-gram score | Baseline Model | Fine-tuned Model |
|--------------------------|-------------------|--------------|--------------|--------------|----------------|------------------|
| HANAN TAHER TRUCKING | 0.75 | 0.83 | 0.67 | 0.58 | 0.91 | 0.94 |
| TRUCKING INC HANAN ATHER | 0.00 | 1.00 | 0.88 | 0.79 | 0.93 | 0.96 |
| ATHER TRUCKING INC | 0.75 | 0.75 | 0.71 | 0.67 | 0.73 | 0.49 |
| GODBOUT TRUCKING INC | 0.58 | 0.58 | 0.50 | 0.46 | 0.58 | 0.35 |
| HANAN ATHER PHARMACY INC | 0.67 | 0.67 | 0.58 | 0.50 | 0.59 | 0.56 |
| ATHER INC | 0.38 | 0.38 | 0.33 | 0.25 | 0.48 | 0.29 |

Traditional approaches, such as the Levenshtein score, are highly sensitive to changes in word order, resulting in significantly lower similarity scores (e.g., "TRUCKING INC HANAN ATHER" scores 0.00). In contrast, N-gram methods provide greater robustness by assessing overlapping substrings, with higher N-values (e.g., 2-gram and 3-gram) showing improved resilience. Notably, fine-tuned BERT-based models outperform all other approaches by capturing deeper semantic relationships, achieving the highest similarity scores (0.94 and 0.96) even when word order is altered.

4. Workflow of the Record Linkage Program

The record linkage program follows a structured, multi-stage workflow to maximize efficiency, scalability, and accuracy. The process consists of three main stages—pre-linkage, linkage, and post-linkage—each comprising modular components that facilitate systematic transformations, comparisons, and quality checks.

4.1 Pre-Linkage Stage: Data Preparation and Standardization

The process begins with the pre-linkage stage, where the source data—including the SBR and any external datasets—are imported. These records undergo standardized text processing to mitigate superficial inconsistencies that could impact similarity scoring. Standardization procedures include converting text to uppercase, removing special characters and accents, trimming excess whitespace, and compressing spaces. This ensures that differences in spelling,

punctuation, or formatting do not unduly influence downstream comparisons. Once the data is standardized, the program partitions records into manageable subsets using a blocking strategy. Blocking reduces computational complexity by grouping records based on predefined attributes such as geographic identifiers (e.g., postal codes, provinces). For instance, the program first searches for matches in small, restrictive blocks like identical postal codes before expanding progressively to larger blocks—such as matching records with the first three digits of the postal code or at the provincial level. To address the trade-off between broader searches and increased false positives, penalties are applied to scores at each successive round.

4.2 Linkage Stage: Candidate Pair Matching and Scoring

Within each block, the linkage stage identifies candidate pairs and calculates similarity scores using multiple algorithms. Character-level methods, such as Levenshtein edit distance, capture minor spelling variations, while n-gram overlap methods compare sets of contiguous substrings to detect textual similarities independent of word order. The modular design allows the program to compute these scores efficiently, assembling a comprehensive profile for each candidate pair.

4.3 Post-Linkage Stage: Quality Assessment and Output

In the **post-linkage stage**, the system evaluates the similarity scores against pre-defined thresholds to classify pairs as matches or non-matches. Matched pairs undergo further quality assessments, which may include cross-validation with supplemental attributes (e.g., phone numbers, industry codes), adjustments to weighting schemes, or manual verification for ambiguous cases. The results are then consolidated into a final output table, which includes matched pairs, their associated scores, and **quality ratings** reflecting confidence levels. These quality indicators enable analysts to interpret linkage reliability and tailor their decisions to the specific requirements of their applications.

Throughout the workflow, careful consideration is given to computational efficiency and methodological transparency. The modular nature of the system facilitates continual refinement, allowing new string-matching algorithms, embedding models, or adaptive blocking strategies to be seamlessly integrated as they emerge. By balancing scalability, precision, and flexibility, this workflow provides a robust foundation for record linkage, capable of evolving alongside advancements in Natural Language Processing and statistical methodologies.

5. Challenges and Lessons Learned

The implementation of the record linkage process revealed several key challenges and lessons learned, which provide valuable insights for future projects. These challenges relate to the selection of blocking variables, the importance of tailored approaches, variations in time requirements, and the balance between match quality and overall match rate.

5.1 Impact of Blocking Variables

Blocking variables play a critical role in determining the computational efficiency of the linkage process. The choice of appropriate blocking variables significantly influences computation time, as they reduce the search space by grouping records into subsets. Poorly chosen blocking variables can lead to either excessive computational costs or missed matches. As such, careful consideration and experimentation with blocking criteria, such as geographic attributes or key identifiers, are necessary to optimize the trade-off between speed and completeness.

5.2 Tailored Approaches for Each Project

There is no one-size-fits-all method for record linkage. Each linkage task requires a customized approach tailored to the specific data sources, project goals, and variables involved. Most customization occurs at the **beginning** of the process—during data preparation and blocking—and at the **end**, where results may require manual review or validation. This flexibility ensures that the methodology aligns with the unique challenges posed by individual projects.

5.3 Variability in Linkage Completion Time

The time required to complete the linkage process varies widely based on factors such as dataset size, complexity, and the number of variables used for comparison. Smaller projects can be completed in under an hour, while large, complex datasets may require several days. Understanding these time variations is crucial for planning and resource allocation, particularly when working with large-scale data integration tasks.

5.4 Balancing Quality and Match Rate

One of the most persistent challenges in record linkage is balancing the trade-off between high-quality matches and the overall match rate. While optimizing for high-quality matches ensures accuracy, it may come at the expense of missing some valid pairs. Conversely, focusing on maximizing the overall match rate increases the risk of false positives. Striking the right balance requires careful calibration of thresholds, weighting schemes, and similarity measures to meet the specific needs of each project.

5.5 Data Quality and Process Optimization

The success of the linkage process is heavily influenced by the quality and completeness of the datasets. Incomplete or inconsistent data can hinder the accuracy and reliability of matches (Archer, 1995), underscoring the importance of robust preprocessing and quality control. Additionally, process optimization—tailoring the methodology to address project-specific challenges—is essential to ensure efficient and accurate linkage. Each project demands iterative refinement of parameters and methods to achieve optimal performance.

6. Future Work and Conclusion

6.1 Future Work

While the present methodology demonstrates improvements in efficiency over earlier name-based linkage procedures, there remain opportunities for further enhancements. One promising avenue lies in deepening the integration of transformer-based embeddings. Although the current approach leverages semantic vector spaces to account for textual complexity, future work may experiment with specialized pretrained models focusing on business and economic corpora. Domain-adapted embeddings could capture industry-specific terms, acronyms, and naming patterns more effectively, thereby increasing both precision and recall.

Another important direction involves refining the underlying algorithms that govern blocking, scoring, and matching decisions. Implementing more adaptive blocking strategies, for example, could dynamically adjust partitioning criteria based on evolving patterns in incoming data, optimizing for efficiency and accuracy as the datasets grow. Incorporating active learning techniques and semi-supervised approaches could further refine threshold selection and quality assessments, reducing reliance on arbitrary cutoffs and strengthening the alignment between the linkage outputs and user-defined standards of quality. In addition, the program's migration to Python and the integration of new libraries and frameworks will expand the range of available NLP tools, ultimately enabling more continuous updates and seamless experimentation.

6.2 Conclusion

In conclusion, the methodology and system described here represent a significant improvement in the ability to accurately and efficiently link external datasets to the SBR by business names. As future enhancements are introduced, the program stands to further improve its accuracy, adaptability, and transparency, thereby strengthening the foundational data infrastructure that supports Canada's economic statistics and research endeavors.

References

- Oyarzun, J., and Wile, L. (2016), “An Overview of Business Record Linkage at Statistics Canada: How to Create an Effective Business Register”, Statistics Canada.
- Reimers, N., and Gurevych, I. (2019), “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, pp. 3982-3992.
- Archer, D. (1995), “Maintenance of Business Registers”, in B. G. Cox et al. (eds.) *Business Survey Methods*, New York: Wiley, pp. 85-100.
- Levenshtein, V. I. (1966), “Binary Codes Capable of Correcting Deletions, Insertions, and Reversals”, *Soviet Physics Doklady*, 10(8), pp. 707-710.