Analysis of Address Check Tool Performance

Team Datascience (Benjamin Berthold)

DENIC eG

February 5, 2025

Contents

1 Introduction

This report provides a brief overview of our toolt's performance in our own benchmarks, focusing on the top 10 country codes plus Germany. The analysis highlights key results, demonstrating the tools effectiveness and reliability in handling different regions globally.

2 Benchmarking Methodology

For benchmarking, we use a dataset of 1,430 manually verified addresses. This dataset is evenly distributed across the 11 country codes in focus, with each country represented by 90 correct addresses and 40 deliberately incorrect ones. The incorrect addresses either contain inaccurate components or are entirely fabricated, while still adhering to the format and conventions of the respective country. This approach ensures a balanced and realistic evaluation of our tools performance.

2.1 Address Source

The addresses were randomly selected from each respective country and manually verified to ensure accuracy. This process guarantees a representative and high-quality dataset for benchmarking, covering both correct and intentionally incorrect addresses in a controlled manner.

2.2 Benchmark Execution

The benchmark evaluates all addresses against the API in the respective version. The results are then analyzed using the following key metrics: Precision, Recall, Accuracy, and F1 Score. These metrics provide a comprehensive assessment of the tools address-matching performance.

2.3 Benchmark Metrics Explained

2.3.1 Precision (Positive Predictive Value)

Definition: Precision measures the proportion of correctly identified addresses (true positives) out of all addresses that the tool classified as correct (true positives + false positives).

Formula:

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}} \tag{1}$$

Interpretation for Address Matching: A high precision score indicates that when the tool classifies an address as correct, it is usually right. A low precision score means the tool frequently misidentifies incorrect addresses as correct (false positives).

2.3.2 Recall (Sensitivity or True Positive Rate)

Definition: Recall measures the proportion of actual correct addresses that the tool successfully identifies.

Formula:

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} \tag{2}$$

Interpretation for Address Matching: A high recall score means the tool successfully detects most of the correct addresses, while a low recall score suggests it frequently fails to recognize valid addresses (false negatives).

2.3.3 Accuracy

Definition: Accuracy measures the overall proportion of correctly classified addresses (both correct and incorrect) in the dataset.

Formula:

$$\text{Accuracy} = \frac{\text{True Positives (TP)} + \text{True Negatives (TN)}}{\text{Total Samples (TP + TN + FP + FN)}} \tag{3}$$

Interpretation for Address Matching: Accuracy provides a general performance indicator but may be misleading in imbalanced datasets (e.g., if correct addresses vastly outnumber incorrect ones).

2.3.4 F1 Score (Harmonic Mean of Precision and Recall)

Definition: The F1 Score balances precision and recall by calculating their harmonic mean, providing a single score that considers both false positives and false negatives.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

Interpretation for Address Matching: A high F1 Score indicates that the tool effectively balances precision and recall, meaning it neither incorrectly classifies too many addresses nor misses too many valid ones.

3 Baseline

For our benchmark, we use version 1.0.9, which is currently deployed in our production environment. This version serves as the baseline for our evaluation, providing a reference point for measuring improvements in future iterations. The following results outline the performance of our tool across different country codes, assessed using key metrics such as Precision, Recall, Accuracy, and F1 Score.

3.1 Precision

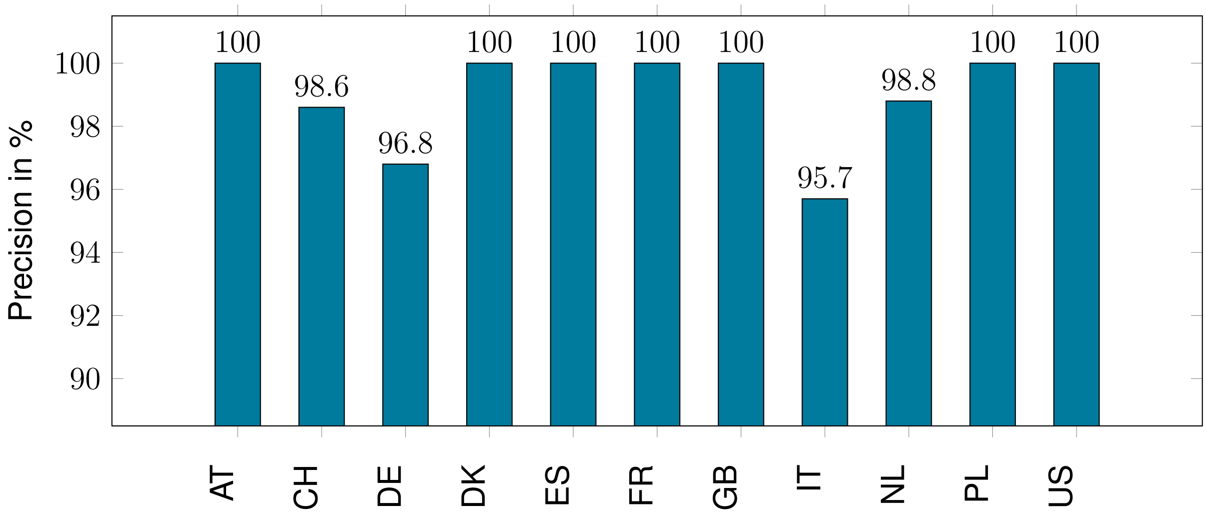

Figure 1: Baseline Precision

3.2 Recall

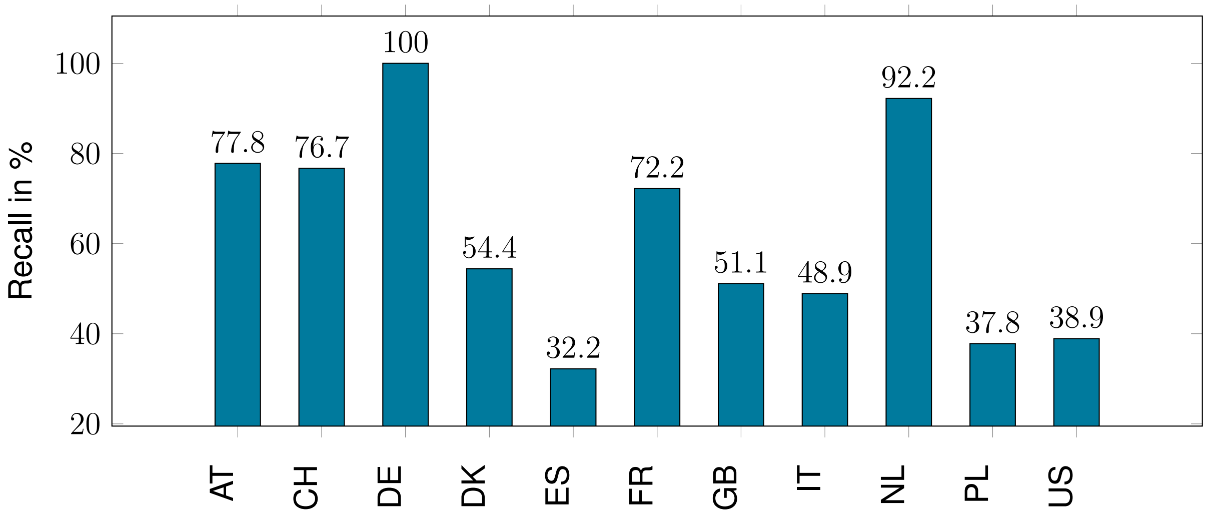

Figure 2: Baseline Recall

3.3 Accuracy

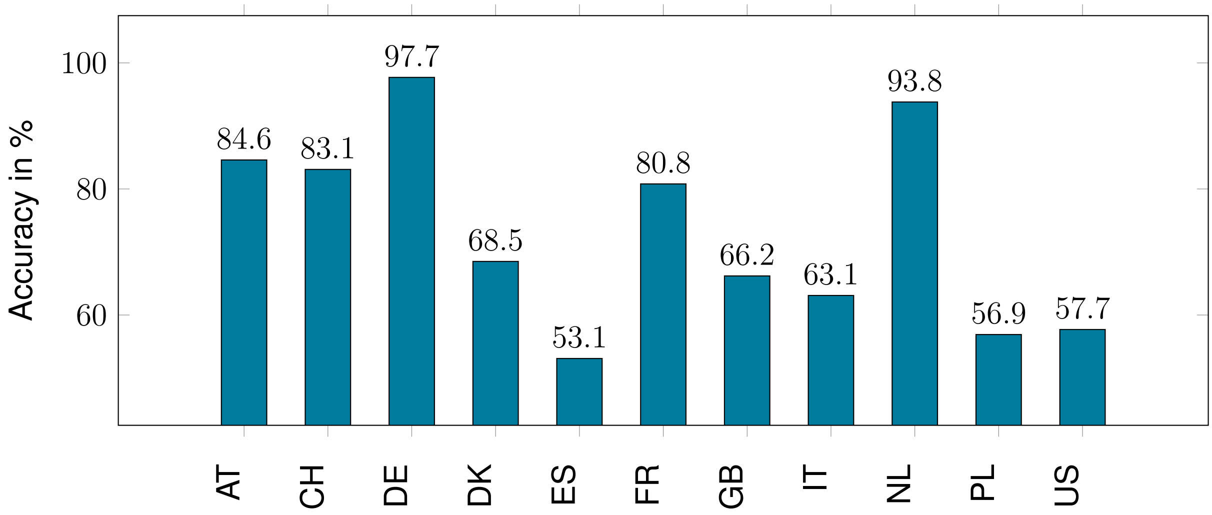

Figure 3: Baseline Accuracy

3.4 F1 Score

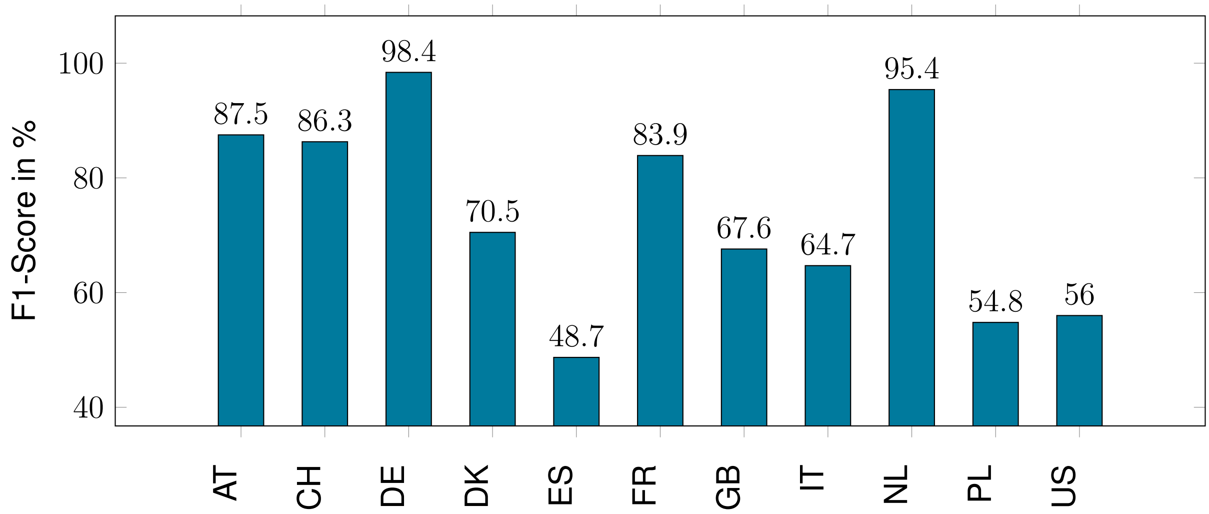

Figure 4: Baseline F1 Scores

4 Improvements

4.1 Danish Addresses

Equipped with the tips and tricks for Danish addresses from *Nikolaj Ravn Hansen*, we have attempted to improve the matching performance by applying country-specific adjustments to the input and output processing.

Source Link: Centr Gitlab Workforce DK Folder

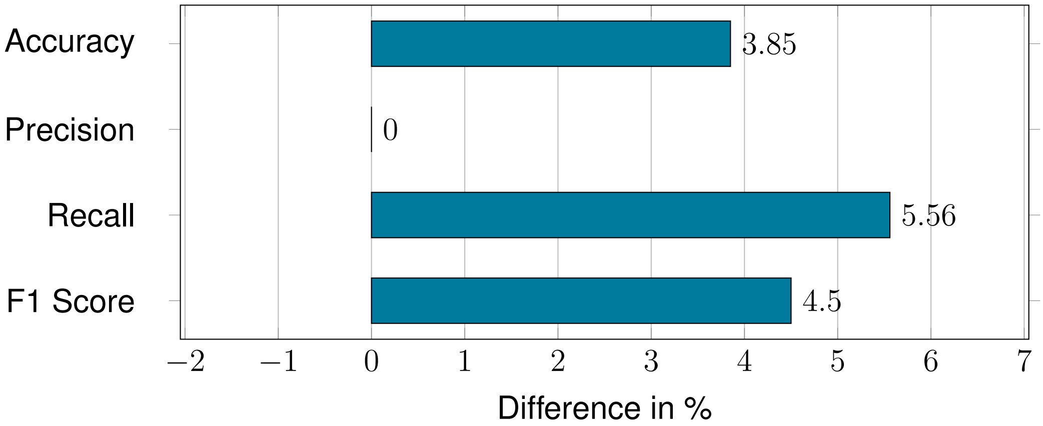

Figure 5: Difference to Baseline Complete Addresses (DK).

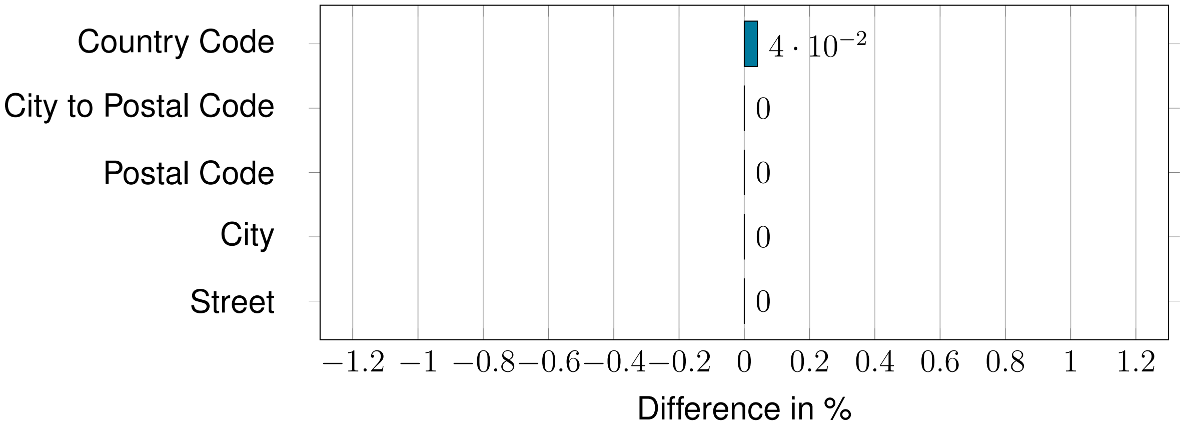

Figure 6: Difference to Baseline Address Parts Precision (DK)

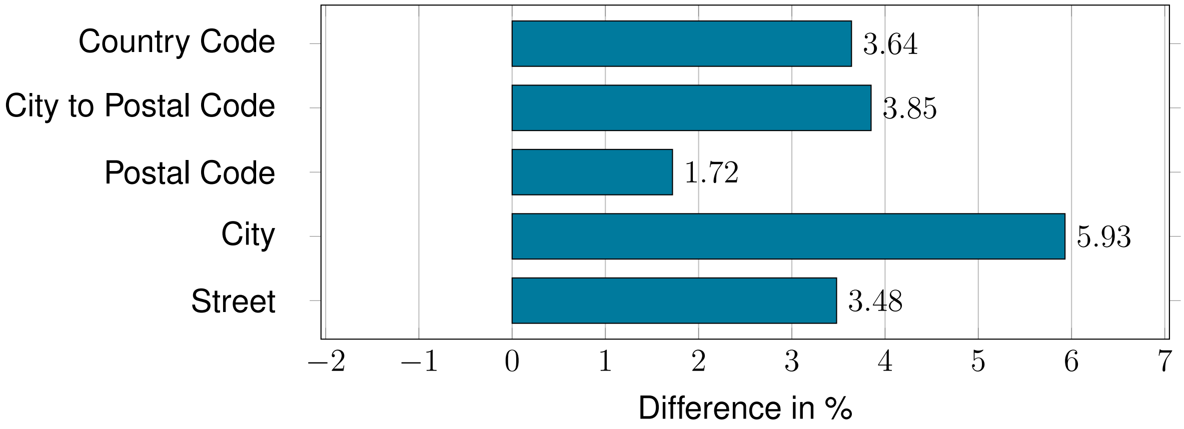

Figure 7: Difference to Baseline Address Parts Recall (DK)

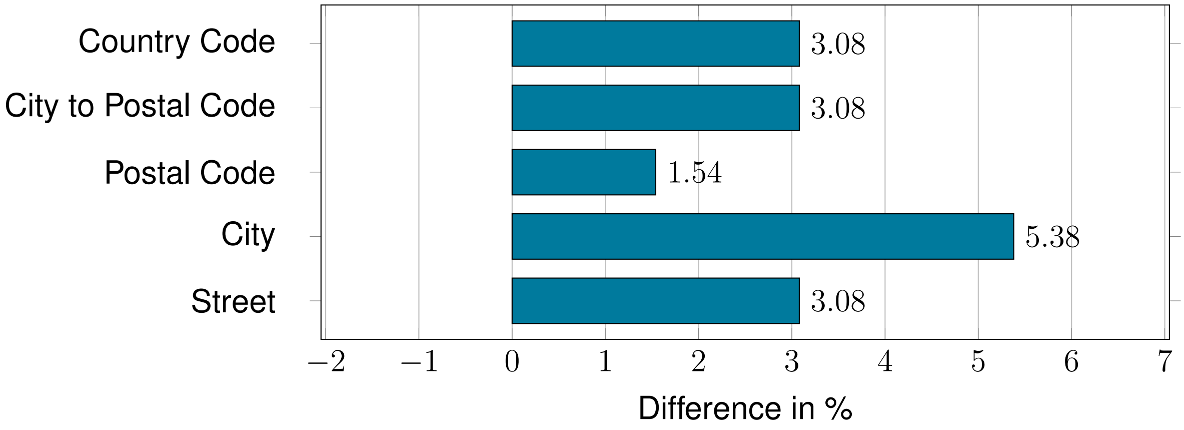

Figure 8: Difference to Baseline Address Parts Accuracy (DK)

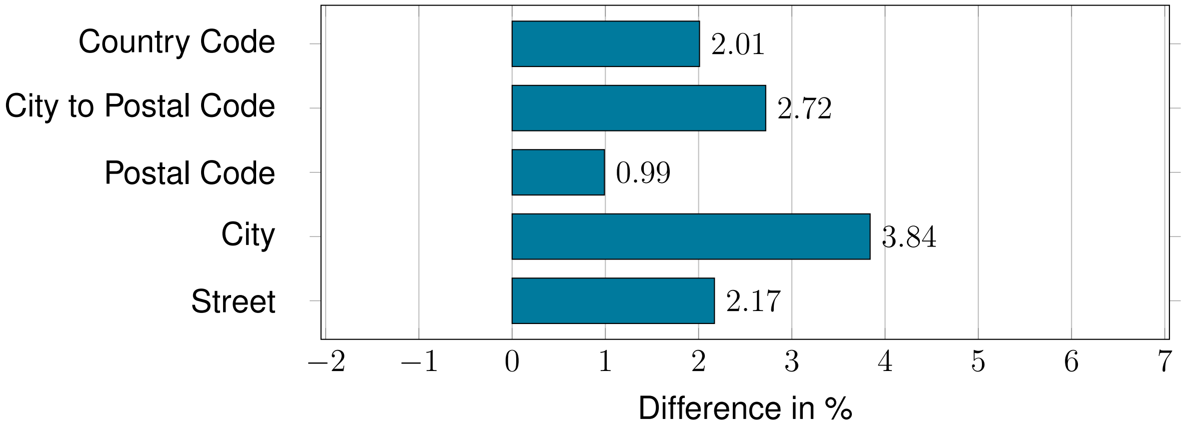

Figure 9: Difference to Baseline Address Parts F1 Score (DK)

Conclusion: The relatively small improvements in the results for Danish addresses can be attributed to the already strong performance in the baseline benchmark. However, real-world scenarios are likely to benefit more significantly from these optimizations.

4.2 US Addresses

The improvements for US addresses were prioritized early on, as the tool's performance for US address matching was relatively poor. By focusing on key weaknesses, we aimed to enhance accuracy and overall reliability in this region.

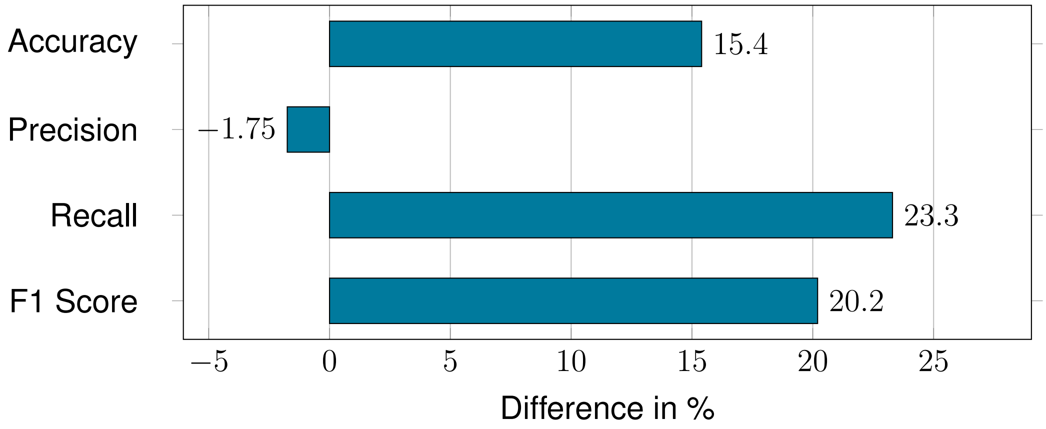

Figure 10: Difference to Baseline Complete Addresses (US).

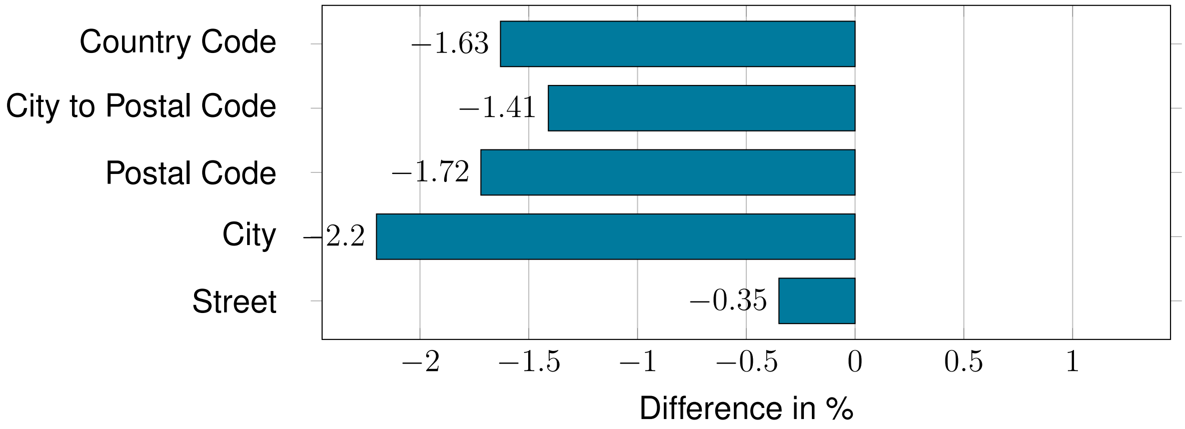

Figure 11: Difference to Baseline Address Parts Precision (US)

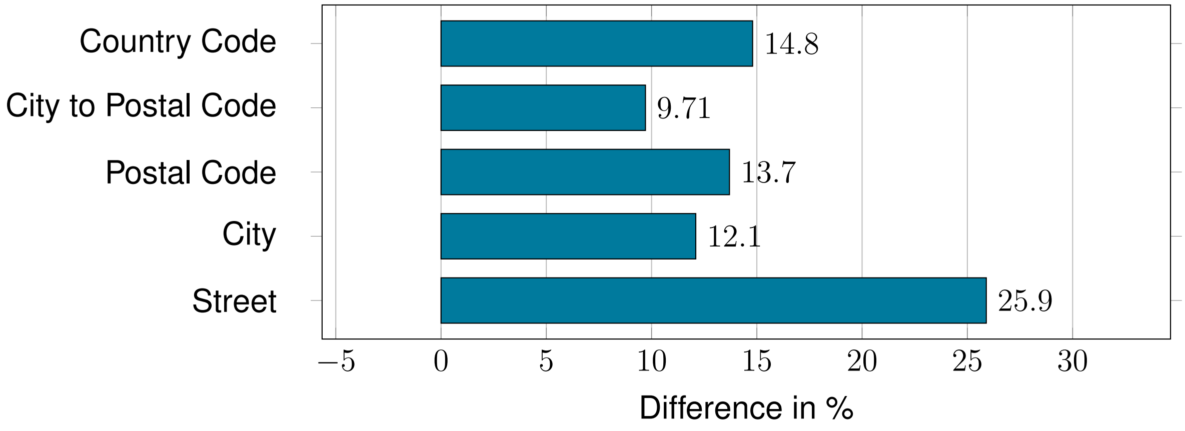

Figure 12: Difference to Baseline Address Parts Recall (US)

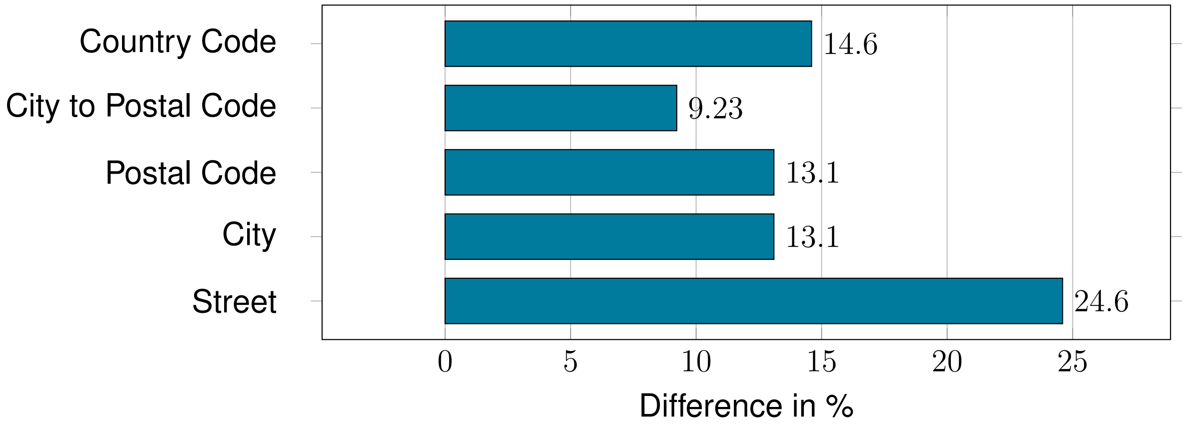

Figure 13: Difference to Baseline Address Parts Accuracy (US)

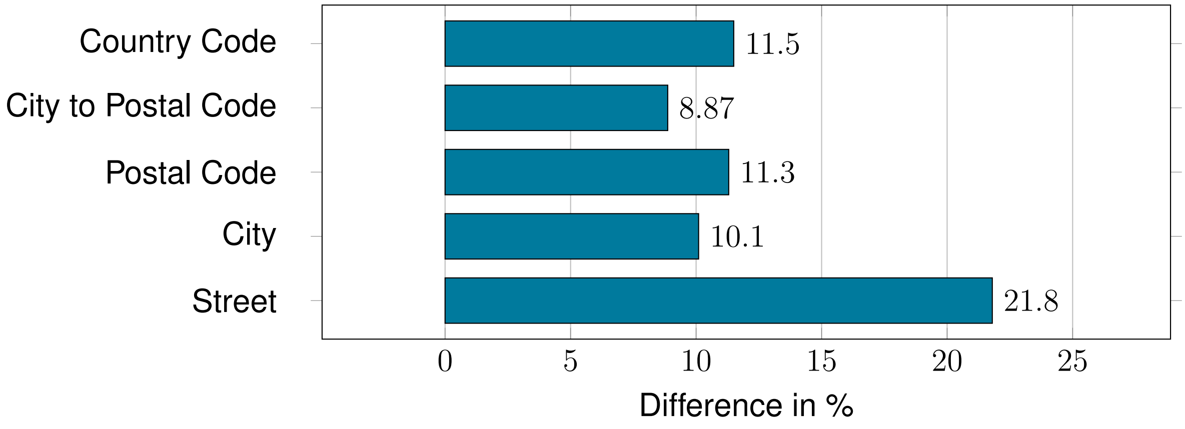

Figure 14: Difference to Baseline Address Parts F1 Score (US)

Conclusion: The adjustments made for US addresses resulted in a significant performance boost, highlighting the fundamental differences between US and European address structures. These optimizations helped bridge the gap, making the tool more reliable for handling US-specific address formats.