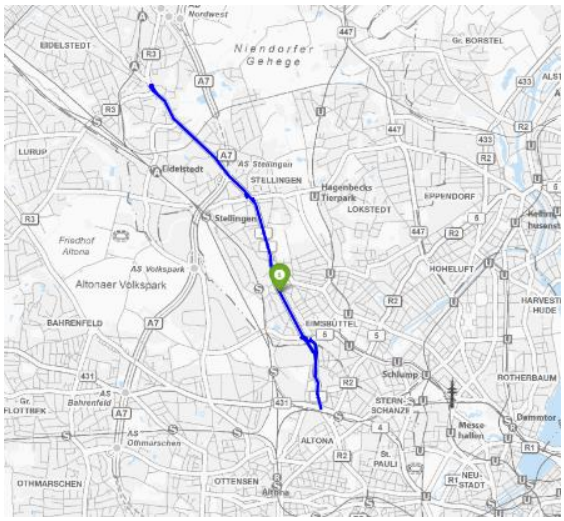
Geodaten aus Texten mit KI extrahieren

Dokumentation des Deep Dives

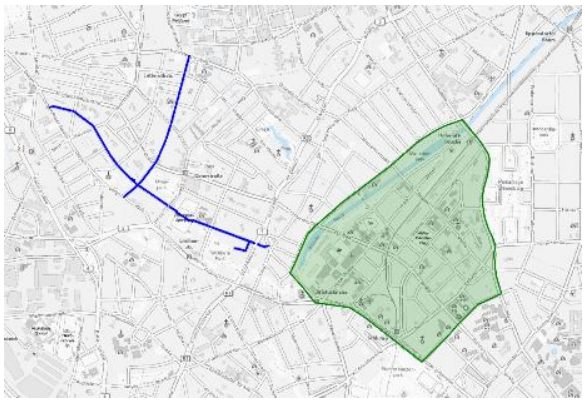
David Engler (Landesbetrieb Geoinformation und Vermessung/Urban Data Analytics)

Abbildung 1: David Engler moderiert den Deep Dive zur Extraktion von Geolocation aus Texten - (C) Angela Pfeiffer

In diesem DeepDive wurden zunächst die in der Bachelorarbeit erarbeiteten Methoden zur Extraktion von Geolokationen aus Texten vorgestellt. Ausgangspunkt ist, dass Nutzer im aktuellen System Geometrien zu den Texten angeben können, diese Angaben aber häufig fehlerhaft sind. Es treten sowohl falsche Geometrietypen auf. In diesem Beispiel „Der gesamte Radweg entlang der Kieler Straße ist nicht befahrbar ...“ wird beispielsweise statt der Straße nur ein Punkt auf der Straße angegeben:

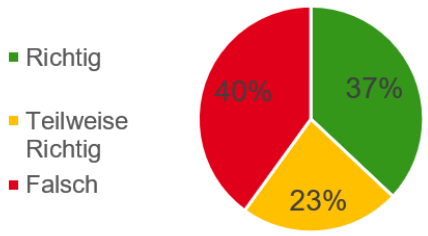


Aber es wurden auch komplett falsche oder falsch verortete Geometrien angegeben. Beispielsweise wurden in folgenden Beispiel zwei Straßen in dem Text beschrieben, aber ein Polygon, neben diesen Straßen angegeben:



Um zu evaluieren, wie häufig diese Diskrepanz auftritt wurde ein Datensatz mit 100 Geometrien untersucht.

Richtig	37
Teilweise Richtig	23
Falsch	40

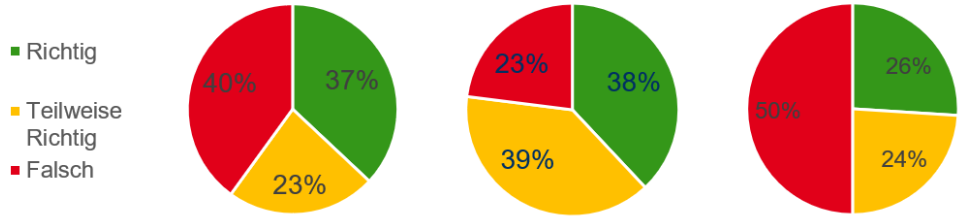


Im Anschluss wurden die entwickelte Methodik zur Extraktion der Geodaten vorgestellt. Zunächst wurde der Regelbasierte Ansatz vorgestellt. Bei diesem Ansatz werden Muster im Text identifiziert, um relative Geometrien zu extrahieren. Dazu gehören die Erkennung benannter Entitäten, eine Rechtschreibkorrektur dieser Entitäten sowie die Klassifikation von Sätzen auf Basis von Schlagwörtern wie „Kreuzung“, „Ecke“ oder „Gabelung“. Mit spezifischen Geodaten, wie dem Straßennetz oder einem Verzeichnis an Parks und Plätzen können anschließend passende Geometrien ermittelt und verarbeitet werden. Im Anschluss wurde der LLM-Ansatz vorgestellt. Hier wurde getestet, inwieweit ein Sprachmodell (LLM)

Beschreibungen von Geometrien interpretieren kann. Text und System-Prompt werden an das Modell übergeben, das daraufhin eine Datenbankabfrage generiert. Mit dieser Abfrage lassen sich die entsprechenden Geometrien zu den Texten auslesen.

Diese Methodik wurde im Anschluss wieder an den 100 Beiträgen evaluiert. Dabei sind folgende Ergebnisse entstanden:

	Originale Geometrien	Geometrien regelbasierter Ansatz	Geometrien LLM-Ansatz
Richtig	37	38	26
Teilweise Richtig	23	39	24
Falsch	40	23	50



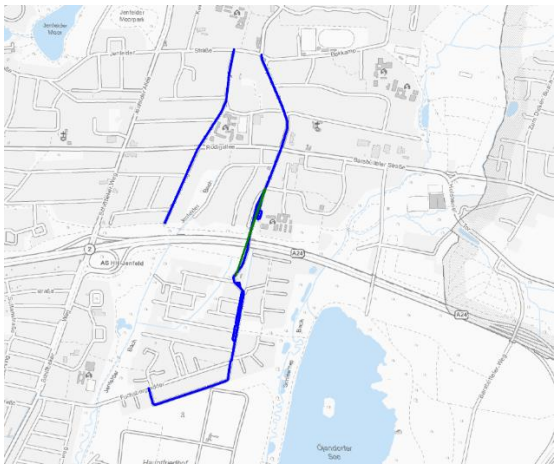
Der regelbasierte Ansatz zeigte gute Resultate, insbesondere bei klar und eindeutig beschriebenen Geometrien. In diesen Fällen konnten die im Text enthaltenen Informationen zuverlässig erkannt und den passenden Geometrien zugeordnet werden.

Der LLM-Ansatz erwies sich als vielversprechend, da er auch bei komplexen oder uneindeutigen Beschreibungen in der Lage war, korrekte Geometrien zu extrahieren. Gleichzeitig traten jedoch häufig Fehler auf, die sich in Form von Halluzinationen oder falschen Interpretationen äußerten.

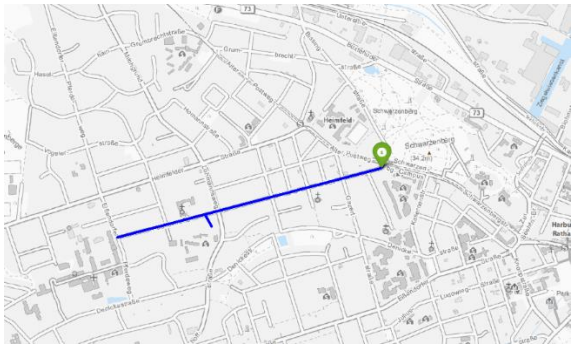
Bei der Bewertung der Ergebnisse ist zudem zu berücksichtigen, dass unter den fehlerhaft extrahierten Geometrien auch solche Fälle enthalten waren, die sich grundsätzlich nur schwer oder gar nicht automatisch aus Texten ableiten lassen. Diese sogenannten Edge Cases umfassen etwa mehrdeutige Formulierungen, unvollständige Angaben oder Veränderungen im Straßennetz.

Edge Cases

- Mehrdeutigkeiten
    - Schulen werden nach Straßennamen benannt -> Modelle extrahieren die Straßen:
      „… den Schulen Öjendorfer Damm und Denksteinweg läuft über …“



- Geometrien werden nicht komplett im Text genannt
    - Kreuzung wird nur anhand der angegebenen Geometrie und einem Straßennamen beschrieben



- Neuerungen im Straßennetz

  „Inzwischen gibt es am Högerdamm …“

  Högerdamm -> Recha-Lübke-Damm, Bella-Spanier-Weg

- Fehler/Verwechselungen der Nutzer

  „Parkplätze an der Großen Bergstraße entfallen lassen ....“

- Außerdem weitere Inputs aus der Runde
    - Mehrfach vergebene Namen (Hauptstraße gibt es in den meisten Orten Deutschlands)
    - Namen werden nur als Vergleich gelistet
    - Planungsvorhaben

- Inoffizielle Namen
- Beschreibungen relativ zu Objekten (… hinter XY, … neben XY)
- Nur indirekt genannte Lokationen (Straße vor der Kirche, …)

Weitere Anwendungsmöglichkeiten aus der Runde

- Für DIPAS
  - Barrierefreiheit
    - Nutzer müssen nicht eine Geometrie angeben, sondern können Beschreiben
    - Aber auch andersrum, Geometrien aus einer Karte werden in Sprache umgewandelt
  - Direktes Feedback nach der Eingabe des Textes, ob die extrahierte Geometrie gemeint ist und sonst bitte zur Eingabe der Geometrie
- Außerhalb von DIPAS
  - Verortung bei Notrufen
  - Verortung von Nachrichten, um Aussagen über bestimmte Regionen zu treffen
  - Verortung von Bebauungsplänen
  - Auswertung von Social Media Beiträgen für verschiedenen Anwendungen (Katastrophenschutz, Bürgerbeteiligung, …)

Abbildung 2: Diskussion in der Gruppe - (C) Angela Pfeiffer

Zusätzliche Anregungen aus der Runde

- Wie kann man den extrahierten Geometrien vertrauen?
  - Manuelle Überprüfung scheint ineffizient, da dann die automatische Extraktion unnötig ist
  - Idee von Kombination beider Ansätze und automatisierten Vergleich aller Geometrien

- Weiteres Modell nachschalten, dass anhand einer Karte und dem Text die Ergebnisse überprüfen kann