EXAMENSARBETE Improving the OpenStreetMap Data Set using Deep Learning
STUDENT Hampus Londögård & Hannah Lindblad
HANDLEDARE Pierre Nugues (LTH)
EXAMINATOR Jacek Malec (LTH)

Förbättra OpenStreetMaps Vägnamnsdata genom Neurala Nätverk

POPULÄRVETENSKAPLIG SAMMANFATTNING **Hampus Londögård & Hannah Lindblad**

Vi demonstrerar i denna artikel en generell tre-stegs-lösning för att förbättra vägnamnsdata i olika länder genom att fylla i saknade namn, rättstava felstavade namn och flagga anomalier.

OpenStreetMap (OSM) är en öppen GIS-databas där det inte finns begränsningar på vem som kan ändra datan. Det leder till att fel förekommer. För att förbättra vägnamnsdata så krävs en robust metod som bemästrar alla de svårigheter som finns i språk. En algoritmisk/regelbaserad lösning är inte tillräcklig då det snabbt blir komplext och för varje nytt språk krävs nya, manuellt skapade regler.

I vår lösning används **neurala nätverk** (NN), NN är en undergrupp av maskininlärning. Detta betyder att NN kan ”lära” sig att lösa en uppgift bättre utifrån mer data utan att explicit bli programmerad för uppgiften. NN försöker efterlikna den mänskliga hjärnan och dess neuroner. Artificiella neuroner kan lära sig att antingen släppa igenom signaler eller inte och de mer avancerade har dessutom ett litet minne med möjlighet att glömma.

Saknade vägnamn i OSM-datan fylls i genom att algoritmiskt hitta 1) om en komponents grannar har samma namn som varandra och 2) om komponenten ingår i en liten ”svans” som går ut och in ur samma nod. Se figur 1.

Anomali-flaggning kombinerar namn och andra tillhörande taggar, exempelvis hastighet och vägunderlag, för att hitta anomalier i datan. En anomali är exempelvis en *gade* med en max-

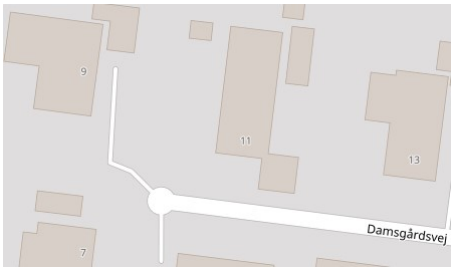

Figure 1: Exempel på ”svans” utan namn.

hastighet på 120 km/h. Vi förbättrade resultatet genom att utnyttja heuristiken att slutet generellt väger tyngst i nordiska språk (tänk: anders*vej*, hjem*stien* och sol*gade*). Med andra ord så utnyttjar vi enbart de fem sista bokstäverna av vägnamnet som indata.

Rättstavaren bygger på **sequence-2-sequence**-struktur vilket är ett **recurrent neural network** (RNN). Denna struktur låter nätverket läsa in och generera sekvenser av obestämd längd, nätverket läser en bokstav i taget och genererar nästa beroende på historiken av lästa bokstäver. För att ta vårt system steget längre så expanderade vi på detta koncept och utvecklade nätverket till ett **bidirectional RNN** vilket betyder att nätverket läser ordet både fram- och bakifrån innan den genererar text.

EXAMENSARBETE Improving the OpenStreetMap Data Set using Deep Learning
STUDENT Hampus Londögård & Hannah Lindblad
HANDLEDARE Pierre Nugues (LTH)
EXAMINATOR Jacek Malec (LTH)

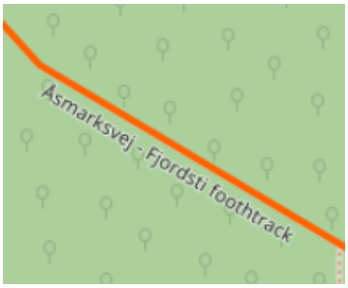

Figure 2: Exempel på felstavat namn.

Detta leder till att ändarna får en större betydelse och att systemet "kan se framtiden", det vill säga bokstäverna framför. Anledning till att vi gör så är att vi ser själva problemet som ett översättningsproblem ifrån fel- till rättstavat.

Resultaten visar först och främst att vår rättstavare är effektivare än en algoritmisk lösning baserad på en riktig ordbok och att Anomaliflaggningen är effektivare än en slumpbaserad lösning med sina 89.3 % F1-poäng och 89.2 % rätt. Saknade namn fyller i 12.5 % av alla saknade namn och görs korrekt i ungefär 70 % av fallen enligt manuell kontroll. Vi visar även ett minst lika bra resultat för Estland som för Danmark genom att träna om systemet på ny OSM-data med estländsk Wiki-data.

Avslutningsvis så kan vårt slutsystem enkelt föras in i **Atlas Checks** som stödjer uppladdning till **MapRoulette** för att bidra till OSM genom att flagga instanser för närmare kontroll.