Opsamling på workshop om store danske sprogressourcer

November 2023

Baggrund

Tirsdag d. 12. september 2023 afholdte Digitaliseringsstyrelsen en workshop om store danske sprogressourcer. På workshoppen deltog omkring 50 personer fra samlet 30 forskellige organisationer, og formålet var at diskutere relevansen af en række eksisterende, men utilgængelige, danske sprogdata.

Udgangspunktet var data fra Det Kongelige Bibliotek, Rigsarkivet, Lex.dk og Aarhus Stadsarkiv. Hver organisation var inviteret til at holde et 15 minutters oplæg omkring størrelse, kvalitet, struktur, format mv. for deres sprogdata. Slides fra præsentationerne er blevet sendt ud til workshoppens deltagere efterfølgende.

På baggrund af præsentationerne blev der i grupper diskuteret to spørgsmål:

1. De pågældende datas værdi for sprogteknologiske formål.
2. Hvad der teknisk skal til for, at disse ”rå” data kan blive til anvendelige sprogressourcer.

Workshoppen blev afholdt som et led i et større arbejde, hvor Digitaliseringsstyrelsen søger at afdække, hvilke data det giver mening at prioritere i en indsats for at tilvejebringe flere danske højkvalitetssprogdata.

I dette korte slidedeck opsummeres først tværgående pointer fra diskussionerne af workshoppens to hovedspørgsmål. Dernæst præsenteres de konkrete handlingspunkter, som Digitaliseringsstyrelsen på baggrund af workshoppen har sat sig for. Til sidst præsenteres gruppernes diskussionsnoter for hvert af de respektive oplæg.

Tværgående pointer fra workshoppen

DIGITALISERINGSSTYRELSEN

Tværgående opmærksomhedspunkter #1

- Der er rigtig meget inden for EU, hvor det er vigtigt, at vi kan dele data. F.eks. Sveriges GPT model, som ikke har haft adgang til danske og norske data. Det kunne gavne, at vi miksede de data på tværs. Så hvis man kan dele data på tværs af Skandinavien vil det være brugbart. Skal det være ”lukket” for Skandinavien/EU eller være offentligt?

- Værdien af data er kontekstafhængig. Små og private virksomheder har en kort tidshorisont, så de gode data bliver værdiløse for dem, når arbejdet overstiger den tidshorisont, de arbejder ud fra.

- De store virksomheder tager data til trods for juridiske barrierer, så måske er det ikke så vigtigt at være bekymret for det? Hvis vi ikke udstiller det, så opstiller vi kun en barriere for os selv – for de andre har alligevel allerede adgang til det.

- Hvordan løser vi samarbejdet mellem dem, der har den lange horisont og dem, der har den korte horisont?

- Kan man lave en slags ”andelskonstruktion”, hvor nogen faciliterer, at nogle godkendte institutioner processerer data for organisationer, der ikke kan, med udgangspunkt i en udarbejdet protokol. Konstruktionen kunne f.eks. indebære, at ejerne af dataene er Danmark, og at nogen står for at administrerer det. Det er altså ikke en organisation der ejer det (f.eks. Det Kongelige Bibliotek), det er Danmark.

Tværgående opmærksomhedspunkter #2

- Hvor let er det at få adgang til data? Og hvor ligger de og hvem giver adgang til hvad?
- Værdien af data for en dansk sprogmodel; OpenAI kan måske tage vores data og bruge det til kommercielle formål – alternativt kan det give værdi at lave en dansk sprogmodel, hvor vi kender til opbygningen og kan føre kontrol med den.
- Der er et arbejde i gang med at udvikle sprogmodeller.
- Data bør også gøres tilgængelige for europæiske projekter eller nordiske projekter – og det skal undersøges, hvordan det kan gøres kontrolleret.
- Hvis man også skal bruge modeller til medicinske formål og i sundhedssektoren kommer der større krav til, hvad modellerne er trænet på og bruges til – så det vil kræve lokale modeller, som vi har kontrol med. Vi kan måske ikke træne ligeså gode modeller som OpenAI, men det er ikke nødvendigt, for de er ikke en konkurrent i den kontekst.
- Danske modeller som er finetunede klarer sig bedre end OpenAI.

Fremadrettede handlingspunkter

DIGITALISERINGSSTYRELSEN

Næste skridt efter workshoppen

På baggrund af workshoppen vil Digitaliseringsstyrelsen kigge ind i fire handlingsforløb, som er listet nedenfor:

- **Dialog med de fire oplægsholdere**
  - Digitaliseringsstyrelsen vil indgå i en dialog med de fireoplægsholdere for at afdække muligheden for at få (dele af) deres sprogdata tilgængeliggjort og udstillet på sprogteknologi.dk.
- **Netværk for sprogressourcer**
  - Med afsæt i workshoppens drøftelser vil Digitaliseringsstyrelsen udtænke et format for et "netværk for danske sprogressourcer", hvor aktører kan udveksle erfaringer med og koordinere udviklingen af nye danske sprogressourcer.
- **Undersøge mulige løsninger på juridiske udfordringer**
  - Digitaliseringsstyrelsen vil undersøge mulige løsninger på de mest gængse udfordringer med ophavsret og personoplysninger i forhold til at tilgængeliggøre værdifulde data, som på nuværende tidspunkt ikke er udstillet frit.
- **Undersøge muligheden for at lave sprogteknologi.dk om til én sprogressource**
  - For at gøre det lettere at anvende de mange sprogdata, som henvises til på sprogteknologi.dk, vil Digitaliseringsstyrelsen undersøge, om der kan laves én samlet sprogressource på baggrund af de data, der henvises til på sprogteknologi.dk.

DIGITALISERINGSSTYRELSEN

Diskussionsoplæg og noter fra workshoppen

Diskussionsoplæg

På workshoppen var deltagerne inddelt i grupper. Hver gruppe skulle forholde sig til ét af de fire oplæg og skulle diskutere de pågældende data med udgangspunkt i deres værdi og deres efterbehandlingsbehov. Nedenfor er de spørgsmål, der satte rammerne for diskussionen.

VÆRDI	EFTERBEHANDLINGSBEHOV
• Hvilken værdi har de forskellige data – og hvorfor? • Hvilke data har den største værdi – og hvorfor? • Har datene bred sprogteknologisk værdi, eller er de kun velegnede til få og specifikke formål? • …?	• Er det allerede muligt at få trukket data ud og arbejde med det – hvordan? • Skal data efterbehandles inden at de kan anvendes til udvikling – hvordan? • Er der et hensyn, der skal tages i forhold til balancen mellem ”konstruktiv” og ”destruktiv” efterbehandling? • Hvor omfattende er det estimerede nødvendige efterbehandlingsarbejde ift. tid, omkostninger og ressourcer? • Hvordan bør data udstilles efter de er trukket ud og evt. er blevet behandlet for at give det bedste overblik? • …?

På de følgende slides fremgår pointer i noteform, som grupperne præsenterede for hvert af de fire oplæg.

Den Kongelige Bibliotek

VÆRDI	EFTERBEHANDLINGSBEHOV
• Dataene har en generel værdi, f.eks. er Netarkivet et datasæt til forskning, som giver anledning til basismodeller, som kan deles. • Der er potentiel værdi i forhold til skandinaviske- eller EU-datasamarbejder. • Man kan lave den nødvendige efterbehandling på en delmængde i forhold til GDPR, som så kan udgives frit. • Man kan opbygge en kultur omkring danske domæner, hvor man tilføjer om de må blive offentliggjort gennem Netarkivet til sprogteknologiske formål. • Høj redundans, som kan laves til lav redundas – internt eller eksternt. • Værdien er, at de tilgængelige. • Diverse data. • Aktualitet de sidste 20 år, begivenhedshøstninger og temporalitet. • Indekseret til fri tekst.	• De-duplication. • Filtrering af dårlig tekst fra. • Få sat teksten sammen/ekstrahere sammenhængende tekst. • Det skal gøres centralt, så alle ikke behøver at gøre det. Og så er der nogen der skal kuratere de datasæt, så folk kan få adgang. • En hel protokol som skal være open source. Alle metadata kan bare åbnes op. • Temporalitet og begivenhedshøstninger kræver penge.

Aarhus Stadsarkiv

VÆRDI	EFTERBEHANDLINGSBEHOV
• Parallelitet i txt og img. • 200.000 sider er en del. • Edge cases er værdifulde! Håndskrevne tekster skal også behandles teknologisk. • Geografiske metadata og dialekt. • Tidsbestemte metadata → historisk sprogbrug.	• Muligt nu. • Godt med cc-licenser. • Må gerne lægges i åbne databaser. • Txt, jpg mv. er fine (almene) formater. • Parallelisering er allerede sket på sætningsniveau.

Lex.dk

VÆRDI	EFTERBEHANDLINGSBEHOV
• Opdeling i resumé har stor værdi ift. finetuning af resumégenerering. • Autoritativ kilde, så den kan være benchmark for ”sandheden”. • Der er gode metadata og kategorisering. • Ingen GDPR udfordringer. • Tekster kan henføres til forfattere.	• Ikke noget behov. • Det vigtigste er at der er en struktur, ikke hvilken struktur.

Rigsarkivet

VÆRDI	EFTERBEHANDLINGSBEHOV
• Kæmpestort datasæt, særligt mange domænespecifikke data, som ikke ellers ligger frit tilgængeligt på nettet. Så det er ny viden for sprogmodellerne. • Godt til både sprogmodellering og til specifikke opgaver, der er egnet til de specifikke domæner. • Der er en masse samtaler i datasættet, både e-mail og chats, så både formelle og uformelle tekster • Det er fra slut 90'erne og frem til i dag, så virkelig brugbart i forhold til ting, der skal gøres i dag.	• Der er behov for anonymisering, pseudonymisering – og hvornår er det for destruktivt for datasættet? Det kan man først vurdere i brugsscenariet. Der skal findes en balance i det. • Det er meget dyrt og meget tidskrævende. • Billede-til-tekst, det vil tage noget tid, men ikke mere end det. • Der skal være noget kontrol med hvilke data, man kan spørge ind til og låne ud, og de data skal være så rå som muligt. • Der skal måske laves en intern anonymisering, så folk kan få lov til at bruge det. Og man skal passe på at det ikke skjuler for mange ting, som ikke er personfølsomme. Så måske et samarbejde om dette. • Man kan arbejde med forskellige grader og metoder afhængig af hvad man har fat i og hvilken opgave der skal løses. • Hvis man kan lave en heftig filtrering, som fjerner alt personhernførbart, så man når ned i en mindre størrelse, men som til gengæld kan udstilles, så kan det give værdi for mange mennesker.

Tak!

Tak til Det Kongelige Bibliotek, Rigsarkivet, Lex.dk og Aarhus Stadsarkiv for at tage sig tiden til at forberede nogle gode præsentationer og være med til nogle spændende diskussioner på workshoppen.

Tak til de mange sprogteknologiske ildsjæle, som deltog og som løbende hjælper Digitaliseringsstyrelsen med at kvalificere arbejdet med at understøtte dansk sprogteknologi.

DIGITALISERINGSSTYRELSEN