Datamanagementplan for

"*The general reasoning benchmark dataset for Danish*"

V1 27.01.2025

Om projektet

"Benchmarkprojektet" er et samarbejde mellem Det Danske Sprog- og Litteraturselskab (DSL) og Center for Sprogteknologi (CST), NorS, KU om udvikling af et fælles dansk benchmark-datasæt til evaluering af store sprogmodeller (LLMs). Formålet med projektet er mere specifikt at udvikle en leksikalsk semantisk "guldstandard" til brug i forskning og evaluering af danske sprogmodeller.

Projektet er bevilget af Carlsbergfondets forskningsinfrastrukturpulje og løber fra 1.2.2024 til 1.2.2026.

Internt på Center for Sprogteknologi, NorS, KU, fungerer

- Bolette S. Pedersen (bspedersen@hum.ku.dk) som PI.
- Sussi Olsen (saolsen@hum.ku.dk) som data-ansvarlig
- Dorte Haltrup Hansen (dorteh@hum.ku.dk) som dataudvikler
- Simon Gray (simongray@hum.ku.dk) som teknisk ansvarlig

Fra Det Danske Sprog- og Litteraturselskab (DSL) deltager

Sanni Nimb, sn@dsl.dk
Nathalie Hau Sørensen, nhs@dsl.dk

Beskrivelse af data:

Datasættet indeholder ikke personfølsomme data, men bygger på data der indgår i enten DanNet, Den Danske Ordbog, Den Danske Begrebsordbog, Det Danske Framenet eller Dansk Sentiment-leksikon.
Data bliver opstillet som en række sprogforståelsesopgaver, der fokuserer på forskellige leksikalsk semantiske aspekter og kan afprøves på diverse LLM'er. De er grupperet efter disse sprogforståelsesopgaver og lagres ift. disse:

- **Inference**
  Leksikalsk inferens ift. begrebsmæssige, taksonomiske egenskaber. Viden hertil uddrages semiautomatisk fra DanNet (https://wordnet.dk/dannet, licens CC-BY-SA 4.0)
- **Entailment**
  Leksikalsk "entailment" ved visse typer verber. Viden hertil uddrages semiautomatisk fra Det Danske FrameNet-leksikon (DSL, ret åben licens, se https://github.com/dsldk/dansk-framenet/blob/master/LICENSE) og inddrager heuristisk viden.
- **Synonymy**
  Synonymi og nærsynonymi af begreber. Viden hertil uddrages semiautomatisk fra Den Danske Begrebsordbog (rettighedshaver: DSL),

- **DanWic**
  Entydiggørelse af ord i kontekst. Viden hertil uddrages semiautomatisk Den Danske Ordbog (rettighedshaver: DSL),
- **Relatedness**
  Semantisk nærhed i et bredere perspektiv end den snævre synonymirelation. Viden hertil inddrages fra Den Danske Begrebsordbog (rettighedshaver: DSL),
- **Metaphors**
  Metaforisk betydning, herunder en forståelse af, hvordan en metaforisk betydning refererer til den bogstavelig betydning. Viden hertil uddrages semiautomatisk fra Den Danske Ordbog (rettighedshaver: DSL),
- **Sentiment**
  Konnotativ betydning (dvs. positiv eller negativ værdiladning). Viden hertil uddrages semiautomatisk fra det danske Dansk Sentiment-leksikon. (https://github.com/dsldk/danish-sentiment-lexicon, licens CC-BY-SA 4.0),

Datamanagement i projektforløbet

Data udtrækkes til fungerende arbejdsfiler, og for at flere ad gangen og fra begge institutioner kan arbejde i datafilerne samtidig, arbejdes der ofte i Google docs/sheets. DSL har backup-plan for disse data. CST lægger alle data, som de arbejder med, på KU's N-drev: N:\HUM-NORS-cst-projekter\benchmark_Carlsberg\data.
Alle færdige data kopieres løbende til github: https://github.com/kuhumcst/danish-semantic-reasoning-benchmark.

Efter projektets afslutning

Efter projektets afslutning lægges data i CLARINs dataarkiv : https://repository.clarin.dk/repository/xmlui/ og forbliver på CST's github-repositorium: https://github.com/kuhumcst/danish-semantic-reasoning-benchmark.

Det forventes desuden at data udstilles gennem sprogteknologi.dk

I projektansøgningen beskrives at data vil blive distribueret med licensen CC BY 4.0. Imidlertid er det sidenhen blevet fastslået at evalueringsdata der ligger frit tilgængelige for download, vil blive brugt som træningsdata for nye store sprogmodeller, hvilket medfører dataforurening og gør data ubrugelige til evaluering, se fx Jacovi et al. 2023. Data distribueres i stedet med licensen CC BY-ND 4.0 og vil kunne downloades sammen med dokumentation af datasættets dækning og tekniske specifikationer. Ved download kræves der desuden et password som fremgår af readme-filen. Dette tjener udelukkende til at forhindre automatisk høstning af data.

Referencer

Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. 2023. Stop Uploading Test Data in Plain Text: Practical Strategies for Mitigating Data Contamination by Evaluation Benchmarks. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 5075–5084, Singapore. Association for Computational Linguistics.