EUROPEAN LANGUAGE DATA SPACE

Language Data in the Age of Large Language Models

Fabrice Nauze (LexisNexis, The Netherlands)
fabrice.nauze@lexisnexis.nl

23-01-2024 LDS Country Workshop, Breda, The Netherlands
https://language-data-space.ec.europa.eu

Inleiding

LexisNexis®

- Data- en analysebedrijf onderdeel van uitgeverij RELX
- Actief op het gebied van
  - Juridische data en patenten
  - Nieuws
  - Bedrijsinformatie
  - Markt- en landeninformatie

- Doel: de toepassing van Nederlandstalige taal- en spraaktechnologie (TST) te stimuleren en te ondersteunen
- Activiteiten
  - Lezingen en events
  - Uitgave van het blad DIXIT

Taal- en Spraaktechnologie

Toepassingen

- Klantenservice – Overheid en Commercieel
- Informatievoorziening
- Gezondheidszorg

TST taken

- Text Generation/Summarisation
- Acoustic Speech Recognition
- Text to Speech Systems
- Dialog Systems / Chat Bots
- Question-Answering Systems
- Named Entity Recognition & Linking
- Relation Extraction
- Sentiment Analysis

Large Language Models
Voordelen

- Foundation models voor taal en spraak in overvloed
- Makkelijk te gebruiken door TST-leken
- Met uitzondelijke prestaties


GPT-4 Passes the Bar Exam

382 Philosophical Transactions of the Royal Society A (2024)

35 Pages • Posted: 15 Mar 2023 • Last revised: 3 Apr 2024

- Hebben we nog language data nodig? Of TST-specialisten…

Large Language Models
Nadelen

- Zwarte dozen
- Kosten
- Kwaliteit - Commerciële standaarden en normen
  - Specifieke vakkennis is niet per se aanwezig of relevant
- Risico's – Hallucinaties
- Juridische status van gegenereerde data

Here's What Happens When Your Lawyer Uses ChatGPT

A lawyer representing a man who sued an airline relied on artificial intelligence to help prepare a court filing. It did not go well.

The Stanford Daily

News • Science & Technology

Stanford misinformation expert admits to ChatGPT 'hallucinations' in court statement

Large Language Models
In de Praktijk

- Kosten
  - Ouderwetse modellen zijn nog nuttig voor bepaalde taken
- Kwaliteit
  - Veilige toepassingen kiezen
    - Samenvatten/Redigeren
  - Modellen zelf trainen en/of fine-tunen
- Risico‘s beperken - Hallucinaties
  - Complexere architectuur (Retrieval Augmented Generation)

Large Language Models Verbeteren

Eigen Modellen Ontwikkelen

- Enorme hoeveelheid data nodig
  - Tekst/opnamen/beelden
  - Complexe juridische vraagstukken ivm eigendomsrecht, auteursrecht, privacy
  - Basis voor een Foundation Model
- Van embedding naar chat-LLM
  - Reinforcement learning from human feedback (RLHF)
  - Datasets om gedrag en beloningsmechanismen van het model te beïnvloeden

Foundation Model Fine-Tunen

- Heel specifieke datasets
  - Opnamen van sprekers van NL als tweede taal, kinderen/ouderen
  - Juridische of gezondheidszorg datasets
  - Evt PII-gevoelig

Large Language Models
Language Data

1. **Kwantiteit**: publiek beschikbare teksten en corpora
   - Auteur- eigendomsrechten, licentie
2. **Kwaliteit**: specifieke datasets
   - Privé eigendom

3. **Toekomst**: persoonlijke data
   - Incorporeren in model of TST toepassing informatie van/over de gebruikers
   - Persoonlijke fine-tuning
   - Streven naar architectuur waar de gebruiker de data heeft, niet het model

Common European Language Data Space

Thank you!

A Common European Language Data Space – funded under contract LC-01936389 with the European Union.

Fabrice Nauze (LexisNexis, The Nederland)
fabrice.nauze@lexisnexis.nl

23-01-2024 LDS Country Workshop, Breda, The Netherlands
https://language-data-space.ec.europa.eu