GPT-2 og språkets dimensjoner: en analyse av mening, makt og ansvar

Tekst oppfattes ofte som noe dypt menneskelig. Romaner, dikt, avisartikler, emailer og poster på sosiale medier – når vi leser får vi innblikk i et annet menneskes sinn. Til og med den tørreste faglitteraturen er menneskelig: kunnskap overføres, slik kunnskap alltid har blitt delt mellom mennesker. Derfor utfordres vår tolkning av tekst når teksten ikke produseres av et menneske. Hvis en maskin produserer tekst som vekker de samme følelsene som menneskeprodusert tekst gjør, hva er det da som vekker disse følelsene?

I vårt arbeid med å utarbeide GPT-2 møtte vi en rekke interessante spørsmål rundt hvordan språk og mening henger sammen. Tekst, i vid forstand, er humaniora sitt forskningsobjekt – så en kunstig intelligens som produserer tekst er spennende å tolke. Teksten vår modell produserer er interessant nok i seg selv, men vi har valgt å se nærmere på spørsmål rundt tekst og mening, språk og makt, og avslutningsvis spørsmålet om hvem som har ansvar for tekst produsert av en modell som vår.

GPT-2 er en kunstig intelligens med oppgave å oversette, svare på spørsmål og generere tekst – oppgaver den mestrer så godt at den ofte ikke kan skilles fra menneske[1]. I vårt prosjekt brukte vi GPT-2 til å svare på spørsmål vi stilte og til å fullføre ufullstendige setninger. Vi trente modellen, som vi hadde tilgang på ferdigtrent, ytterligere på egenvalgte datasett om temaer vi ønsket at den skulle forbedres på. Dermed fikk vi mulighet (og makt) til å trene opp modellen i en bestemt retning: modellen vil nemlig, etter treningen, generere tekst med lik oppbygning og ordforståelse som datasettene den er trent på. Vår modell ble, etter å ha blitt trent opp på datasett om Trump, spesielt god på USAs 45. president og amerikansk politikk – men i hvilken grad produserte den *meningsfull* tekst?

Tekst og mening

Dersom vi skal diskutere hvorvidt vår GPT-2 produserer meningsfull (eller meningsløs) tekst, så er det nyttig å diskutere hva det egentlig betyr at tekst har mening. Innenfor meta-etikk finnes det to ulike syn på språk, kognitivisme og anti-kognitivisme. Førstnevnte holder at setninger uttrykker sanne proposisjoner (i.e. uttalelser) om verden. Så, setningen «det er riktig å utføre gode handlinger», formidler at det finnes handlinger som er «gode». Anti-

[1] Ambalina, «This Entire Article Was Written by an AI (Open AI GPT2)».

kognitivismen, på sin side, holder at setninger *ikke* uttrykker sanne proposisjoner om verden. Setningene er, i kraft av seg selv, verken sanne eller usanne. Innenfor anti-kognitivismen, så har setninger kun mening, dersom mennesker tillegger ordene i setningen mening. «Gode handlinger», refererer ikke til et bestemt sett av handlinger i verden som alltid er «gode», men heller til de handlingene som den som tolker setningen tenker på som gode. I så fall, så er det som er «godt» verken satt eller endelig, men heller diskutabelt og foranderlig.

Dersom vi legger det anti-kognitivistiske synet på språk til grunn, så står GPT-2 igjen på en noe utrygg grunn når det kommer til meningsfull tekst. Et menneske som skriver en setning, skriver med intensjon og tillegger ordene mening. Den uttrykker ikke noe sant eller usant om verden, men den har blitt tillagt mening fra første stund. Kan GPT-2 tillegge ord mening på samme måte? For å tillegge ord mening trengs intensjon, noe en kunstig intelligens neppe har. I så fall har GPT-2 produsert tekst, men tekst som ikke har mening før et menneske leser teksten og tillegger den mening. Da er det tvilsomt hvorvidt tekstproduksjonen til GPT-2 egentlig holder mål.

Språk og makt

Et annet interessant aspekt av språkproduksjonen til GPT-2 er maktdimensjonen. Språk er makt: før begrepet "seksuell trakassering" ble lansert av Lin Farley på 70-tallet[2] var det flere kvinner som sluttet i jobbene sine, siden de ikke hadde språket til å rapportere om trakasseringen de opplevde. Språk skaper også konseptuelle landskap: innenfor etikk har Alison Jaggar argumentert for at tradisjonell etikk mangler terminologi rundt familierelasjoner, omsorg of følelser - noe som gjør at kvinner ikke har kunnet diskutere problemer relevant for dem, innenfor allerede etablerte etiske rammeverk[3].

Som nevnt innledningsvis, produserer GPT-2 ved at den blir matet med datasett. Dermed er det begrenset hvilke begreper den kan benytte seg av og hvilke konsepter den har tilgang på. Er dette et problem? Kanskje ikke. Kanskje det ikke er opp til maskinene og komme opp med det neste begrepet som frigjør undertrykte grupper i samfunnet. Kanskje det ikke er nødvendig med maskiner som dekker både menn og kvinner sine opplevelser likeverdig, eller venstre-side politikk og høyre-side politikk for den saks skyld. Kanskje disse konsekvensene er "verdt det", men, det kommer til å gå på bekostning av undertrykte og marginaliserte

[2] Farley, *Sexual Shakedown* (New York: Warner Books, 1980).
[3] Jaggar, "Feminism in ethics – Moral justification", 244.

grupper: de som allerede opplever at språket som brukes ikke dekker sine opplevelser vil mest sannsynlig yttligere føle det når data-modeller som har et begrenset språk brukes.

I forhold til vårt prosjekt så handler denne maktdynamikken om hvilke datasett vi ga teksten vår. Amerikas 45. president er en kontroversiell figur, som det er vanskelig å behandle nøytralt. Flesteparten av datasettene vi valgte ut at var fra venstreorienterte avisen New York Times, som var i opposisjon til Trump. Dette påvirket GPT2-en vår til å bli mer negativ mot Trump, enn den hadde blitt dersom vi kodet den på presidentens egne tweets eller den høyrepopulistiske nyhetskilden Fox News. For å unngå for mye bias la vi også inn mer objektive datasett: statistikk over amerikaneres støtte til presidenten og utdrag fra debatter mellom politikere. Det ligger altså makt i å velge datasett, men hos hvem ligger dette ansvaret?

Ansvar

Et vanlig etisk spørsmål som gjelder kunstig intelligens handler om hvem som skal stå ansvarlig for outputen den produserer. Mennesker kan stilles ansvarlig for utsagn og tekster de har produsert, men hvem, eller eventuelt hva, har ansvar for den genererte teksten? Er det de som har utviklet modellen eller de som tar den i bruk?

La oss se for oss at det er ønskelig med et verktøy som kan gi svar på alle typer spørsmål. Vi kan da tenke oss at utviklerne lager modellen med ønske om at den skal produsere god tekst som enkelt forklarer ulike temaer. Da blir kunnskapen verktøyet formidler mer tilgjengelig. Intensjonen til utviklerne er dermed god, og de ville trent modellen på så mange og nøytrale, i den grad kilder kan være nøytrale, datasett de hadde tid og ressurser til. Siden målet er at den skal kunne ha informasjon om alle typer tema, vil det derfor være en dybdebegrensing i datasettene den blir trent på. Med dette mener vi at den ikke har tilgang på å få dyp kunnskap om alle temaer. Det vil derfor være nødvendig at den trenes videre når den blir tatt i bruk av andre, når treningen på temaet de ønsker å bruke modellen til, ikke vil være bra nok.

I den ytterligere treningen kan personer med onde intensjoner slippe til og trene modellen på datasett som inneholder hatefulle ytringer og farlige holdninger. Dette vil komme frem i teksten modellen produserer etter treningen. Teksten som da produseres vil dermed være potensielt manipulerende ovenfor lesere. Burde utviklerne stilles ansvarlige for at de har

produsert en modell som har potensiale til å manipulere et helt samfunn, selv om deres originale intensjon er god?

Dersom svaret på dette spørsmålet er ja, er det trolig ingen som har lyst til å produsere en slik modell. Ingen vil stå ansvarlig for en modell som kan justeres til noe farlig når andre får lov til å trene den. Siden vi ønsker at modeller som GPT-2 utvikles, er det dermed fristende å frigjøre utviklere dette tunge ansvaret – det virker nesten urimelig å skulle stille de til ansvar: er Johann Gutenberg ansvarlig for alle uønskede bøker som har blitt trykket? Samtidig, dersom utviklerne ikke har noe som helst ansvar kan de selv trene den godt og lenge på "dårlige" datasett uten å bli ansvarliggjort for det. På denne måten vil det bli svært vanskelig for brukere å rette opp i biasen maskinen har fått ved utvikling – det vil ta masse tid og ressurser. Så, hvem burde ansvarliggjøres?

Det er klart at utviklere ikke står alene med alt ansvar, men en form for ansvar må de ha: en mulighet kan være å utvikle en form for etisk rammeverk lignende Vær Varsom-plakaten som journalister forholder seg til i sitt arbeid. Brukere av modellen må også ansvarliggjøres, men hvorvidt ansvaret ligger hos brukeren selv eller i lovverket eller hos staten må avgjøres – best vil det nok være dersom det finnes en internasjonal standard på akkurat det.

Denne teksten har diskutert GPT-2 med fokus på mening, makt og ansvar. Hva mening angår er det vanskelig å se for seg tekst produsert av en modell som GPT-2 som meningsbærende i seg selv. Teksten er heller meningsbærende i kraft av menneskene som leser den. I forhold til makt så har teksten illustrert hvordan diskusjon forutsetter konseptuelt rom. Dette må utviklere av GPT-2 ha i mente i sitt arbeid. Til slutt har teksten foreslått et delt ansvar for tekst produsert av GPT-2.

Kilder

Ambalina, Limarc. «This Entire Article Was Written by an AI (Open AI GPT2)». *Lionbridge.* 11. november 2019. https://lionbridge.ai/articles/this-entire-article-was-written-by-an-ai-open-ai-gpt2/

Farley, Lin. *Sexual Shakedown: the Sexual Harassment of Women on the Job.* New York: Warner Books, 1980.

Jaggar, Alison. "Feminism in ethics – Moral justification". I *The Cambridge Companion to Feminism in Philosophy,* redigert av Miranda Fricker og Jennifer Hornsby, 225-244. Cambridge: Cambridge University Press, 2000.