

Ja ihminen loi älyn podcast: jakso5, litteraatti

Miten kotimaista tekoälyä kasvatetaan – Sampo Pyysalo

(Eric) Olemmeko me mielestäsi luoneet älyä vai vaan tehokkaan tavan jäljitellä jotain älykkään kuuloista tekstirakennetta?

(Sampo) No tämä on juuri se kysymys, jota yritän yleensä väistää. Me olemme tehneet omaa pientä osaamme siinä, että suomen kielelle olisi näitä menetelmiä olemassa, ja että tämä kieli ei jäisi ihan täysin sivuun. Uskon, että on erittäin tärkeää digitaalisen suvereniteetin perspektiivistä, että se, mitä me teemme, on täysin avointa. Nykyään minun roolini näyttää lähinnä olevan kiroilla supertietokoneille.

(Arno) Tervetuloa Aalto-yliopiston Ja ihminen loi älyn -podcastin pariin. Nimeni on Arno Solin, ja toimin Aalto-yliopistossa koneoppimisen professorina ja olen myös Suomen ELLIS-instituutissa. Kanssani tätä podcastia juontaa Eric Malmi.

(Eric) Toimin vierailevana professorina Aalto-yliopistossa ja Suomen ELLIS-instituutissa. Tänään meillä on vieraana Sampo Pyysalo Turun yliopistosta. Sampo on pitkän linjan kieliteknologian tutkija ja yksi TurkuNLP-ryhmän vetäjistä. Hän on tutkinut sitä, miten koneet käsittelevät kieltä jo kauan ennen nykyistä kielimallibuumia ja ollut johtavassa roolissa suomalaisten kielimallien, kuten FinBERTin ja Silo AI:n kanssa kehitettyjen Poro-mallien, kehittämisessä. Viime vuosina Sampon tutkimus on keskittynyt yhä enemmän suuriin kielimalleihin. Mistä niiden koulutusdata tulee, miten niitä esikoulutetaan, ja mitä sitten, kun jo kaikki käytössä oleva data on käytetty koulutuksen? Lämpimästi tervetuloa Sampo.

(Sampo) Kiitos.

(Arno) Jos minä mietin, että kuka Suomessa on parhaiten vihkiytynyt isojen kielimallien koulutukseen, kaikkiin niihin salaisiin resepteihin ja muuhun, niin varsinkin esikoulutuksen puolella kääntyisin Sampo sinun puoleesi, koska tuntuu että sinä olet se suurin guru tässä. Eric on sitten taas toisaalta jälkikoulutuksen puolella. Joten katsotaan millainen battle tästä tulee esi- ja jälkikoulutuksen suhteen tänään.

(Sampo) Toivottavasti ei jouduta asettamaan ihmisiä mihinkään järjestykseen tässä. Luulen, että Suomessa on aika vähän ihmisiä, jotka ovat erehtyneet yrittämään. Mutta minä nyt satun olemaan yksi niistä, jotka ovat koettaneet saada tällaisia malleja aikaan ihan alkutekijöistä lähtien. Jonkinlaisia onnistumisia on ollut myös mukana matkalla.

(Arno) Monelle tekoäly jotenkin henkilöityy näihin suuriin kielimalleihin, kuten ChatGPT, Gemini, Claude. Maailma on pullollaan kielimalleja nykyään. Mutta voitko sinä Sampo selittää tällaisen suuren kielimallin periaatteen ihmiselle, joka ei välttämättä tiedä neuroverkoista yhtään mitään?

(Sampo) Jos ajatellaan ihan yleisluontoisesti kielimallia asettaen neuroverkot hetkeksi sivuun, niin tällainen generatiivinen kielimalli, kuten nämä GPT:t ja muut jotka mainitsit,

niin voidaan yksinkertaistaen katsoa, että se on malli, joka antaen jonkin joukon edellisiä sanoja ennustaa todennäköisyysjakauman. Tai vielä yksinkertaisemmin, mikä on todennäköisin seuraava sana. Jos esimerkiksi edelliset sanat ovat "Suomen pääkaupunki on", niin kielimalli voi suuresta tekstiaineistosta oppia, että kohtuullisen usein siellä näkyy sanajoukko "Suomen pääkaupunki on Helsinki". Ja ehkä vähän harvemmin "Suomen pääkaupunki on Turku". Ja tästä voidaan ihan yksinkertaisten laskennan ja todennäköisyysmenetelmien avulla laskea todennäköisyysjakauma, että millä todennäköisyydellä mallin pitäisi sanoa "Helsinki", kun edelliset sanat ovat "Suomen pääkaupunki on". Ja millä todennäköisyydellä sen pitäisi sanoa "Turku". Tällaiset puhtaasti sanajonojen esiintymisen laskentaan pohjautuvat mallit olivat vuosikymmeniä käytössä kieliteknologian alalla ja edustivat vielä muutama vuosikymmen takaperin parasta, mihin pystyttiin tällaisessa seuraavan sanan ennustamisessa.

Ja aika iso läpimurto, jossa juuri nämä neuroverkot sitten tulivat mukaan kuvioon, oli kun opittiin ymmärtämään, että sen sijaan että laskettaisiin näitä sanoja, niin aletaan kouluttaa mallia, joka ennustaa tätä jakaumaa tai tätä seuraavaa sanaa. Eli nykyään otetaan neuroverkko, jota koulutetaan olennaisesti luokittelijana. Eli samalla tavalla, kuin yksinkertaisemmassa sovelluksessa voitaisiin kouluttaa neuroverkkoa sanomaan, että tässä sinulle on sähköpostiviesti, onko tämä spämmiä vai ei spämmiä, yksi vai nolla, niin koulutetaan neuroverkko, jolle annetaan syötteenä, että tässä on edellisten sanojen jono ja sitten seuraava sana on Helsinki, Turku, kaunis, suuri. Kaikki suomen kielen sanat ovat ikään kuin näitä luokkia, joita se malli voi ennustaa. Ja sitten koulutetaan sitä luokittelijana, joka sen sijaan, että se ennustaisi kahta luokkaa, spämmiä vai ei spämmiä tai vastaavaa, ennustaa kymmeniä tai satoja tuhansia luokkia, joista jokainen edustaa yhtä tällaisen sanakirjan sanaa.

Aika suuri osa tätä läpimurtoa, jonka takia kielimallit ovat nykyään näinkin kykeneviä, eivätkä enää näitä malleja, jotka ennustivat seuraavaa sanaa vanhoissa puhelimissa, on että nämä neuroverkot, joita koulutetaan nykyään, ovat erittäin suuria. Kymmeniä miljardeja, satoja miljardeja parametreja. Tämä data, jolla nämä mallit koulutetaan, on erittäin laajaa. Kymmeniä biljoonia sanoja. Ja laskenta tämän mallin kouluttamiseen on erittäin suurta. Puhutaan miljoonista tai kymmenistä miljoonista GPU-tunneista. Osoittautui, että kun nämä komponentit laitetaan oikealla tavalla yhteen, niin kielimalli alkaa oppia yhtä sun toista. Eli ei ainoastaan sitä, mikä on todennäköisesti Suomen pääkaupunki tällä hetkellä, vaan kaikenlaisia muita kapasiteetteja, joita siinä koulutusdatassa esiintyy. Esimerkiksi aritmetiikkaa, tietynlaista päättelyä ja niin edespäin.

(Arno) Tuntuu suorastaan ihmeelliseltä tätä taustaa vasten peilaten, että tästä, että "Suomen pääkaupunki on Helsinki", päästään siihen, että nämä yksinkertaiseen tehtävään kootut mallit yhtäkkiä alkavat ymmärtää maailmaa ja tuottaa tekstiä, tarinoita, raportteja, kaikkea. Miten on mahdollista, että ne yleistyvät tämmöiseen?

(Sampo) Joo, ja tulee jossain määrin yllätyksenä tai tuli aikoinaan jossain määrin yllätyksenä myös minulle. Kieliteknologian tutkimuksen parissa jonkin aikaa toimineena, niin ei kyllä monelle tullut varmaan mieleen, että kielimallit voisivat ikinä

olla paljoo enempää kuin leluja jossain mielessä. Tai vähintäänkin pieni komponentti jossain järjestelmässä, josta voitaisiin alkaa puhua tekoälyjärjestelmänä. Eli tämä ajatus, että jo pelkkä kielimalli itsessään voisi olla koko tekoälyjärjestelmä, on melko uusi. Ja uskon, että tuli aika monelle vähän yllätyksenä. Edelleenkin välillä yllätyn siitä, mihin kaikkeen nämä mallit pystyvät. Aika suuri osahan tätä vastausta on, että tätä pystyttiin skaalaamaan ihan tolkkottomasti. Eli sen sijaan, että koulutettaisiin malleja, joilla on joitain kymmeniä tai satoja tuhansia parametreja, niin koulutetaan malleja, joilla on satoja miljardeja. Ja sen sijaan, että koulutettaisiin pienellä aineistolla, niin kuin ohjatussa oppimisessa vielä vuosikymmen takaperin oli peruslähtökohtana, että koulutetaan ehkä tuhansilla tai kymmenillä tuhansilla esimerkeillä, niin koulutetaan biljoonilla. Ja tämä skaala toi mukanaan paljon, paljon yllättäviä asioita. Mutta on tässä yhteydessä mainittava, että skaala yksin ei ole tuonut meitä tähän pisteeseen, jossa me nykyään olemme. Eli pelkän puhtaan kielimallin, tällaisen pohjamallin, skaalaamisen rajat alkavat jo näkyä, ja nykyäänhän sitten tuodaan paljon enemmän erikoistunutta dataa, työkalujen käyttöä sun muuta mukaan, jotta saadaan kielimalleista irti se, mitä näistä johtavista malleista tänään saadaan.

(Eric) Joskus sanotaan hieman runollisesti, että näitä malleja ei rakenneta, vaan ne kasvatetaan. On yleisesti tiedossa pääpiirteittäin, mikä se resepti näiden kasvattamiseen tai ehkä tarkemmin kouluttamiseen on. Mutta minkä takia on vaikeaa kehittää hyviä kielimalleja?

(Sampo) Joo, tosiaan se perusteknologia on onneksi aika hyvin tiedossa. Se osa, jota oikeasti rakennetaan koodissa. Ja se kasvattaminen, se on siis sitten tämä koulutusprosessi. Siinä nämä reseptit ovat vähitellen siirtyneet enemmän ja enemmän salaisiksi. Siis tämä perusajatus, joka oli GPT-1, -2 ja -3, että otetaan internet-dataa, annetaan se supertietokoneelle ja sitten muutaman viikon tai kuukauden päästä sinulla on johtava kielimalli, niin tämän päälle on alettu tuoda aika paljon kaikenlaista muuta jo. Ja kyllä aika suuri osa siitä, jos tässä nyt on jotain salattua tai erityistä mukana, niin liittyy niihin datan ominaisuuksiin. Eli dataan ollaan tuotu ikään kuin luonnollisesti ilmenevien tekstien lisäksi erityisesti tähän koulutustarkoitukseen kehitettyjä aineistoja, joissa esimerkiksi esiintyy dialogirakennetta tai kysymyksiä ja vastauksia. Tai sitten tällaisia ajatusketjuja, joita käyttävät nyt nämä ns. ajattelevat mallit, jotka ovat yksiä näistä viimeisimmistä. Ja tietysti sitten jos mallin halutaan käyttävän ulkoisia työkaluja, verkkohakua, laskuria, ohjelmointikieltä, mitä ikinä, niin sitten pitää kouluttaa datalla, jossa esiintyy tällaisten työkalujen kutsuja ja niiden vastauksia. Eli datan kautta rakennetaan malliin ne kapasiteetit, joita siltä toivotaan.

(Eric) Aivan. Tästä datasta me pääsemme puhumaan myöhemmin vielä lisää. Mutta jos hetki mietitään vielä sitä itse koulutusreseptiä. Arno viittasi, että yleensä voidaan karkeasti jakaa esikoulutusvaiheeseen, pre-trainingiin, ja jälkikoulutusvaiheeseen, post-trainingiin. Miten sinä kuvailisit, mikä näiden kahden eri vaiheen rooli on siinä, millainen siitä mallista tulee tai mitä kyvykkyyksiä sillä mallilla on?

(Sampo) Joo, nämä käsitteet alkavat olla aika liukuvia. Yhä enemmän sellaista dataa, jonka oltaisiin vielä muutama vuosi sitten katsottu kuuluvan vain jälkikoulutukseen, niin aletaan tuoda myös tähän esikoulutukseen. Koska ollaan havaittu, että mitä

aikaisemmin nämä mallit näkevät tiettyä, esimerkiksi ajatusketjuja ja päättelyä edustavaa dataa, niin sitä paremmin ne lopulta omaksuvat niitä kykyjä, joita sen datan kautta halutaan opettaa. Mutta jos nyt otetaan tällainen hieman vanhakantaisempi perspektiivi, niin voitaisiin yksinkertaistaen sanoa, että tämä ensimmäinen pre-training-vaihe tuottaa pohjamallin, joka pystyy vain ikään kuin hallusinoimaan internet-tekstiä jossain mielessä. Eli tuottamaan sen tyyppistä tekstiä, jota se on nähnyt koulutuksessa, joka on pitkälti ollut internet-aineistoa. Eli jos tällaiselle pohjamallille annetaan verkkosivun ensimmäiset sanat, niin se jatkaa ikään kuin samasta teemasta sen näköisesti, mitä verkkosivu olisi voinut olla. Ja sitten post-trainingissä, siis jälkikoulutusvaiheessa, oltaisiin otettu tämä pohjamalli, ja merkittävästi pienemmällä määrällä merkittävästi korkealaatuisempaa ja erikoistuneempaa dataa opetettu malli esimerkiksi vastaamaan kysymyksiin tai toimimaan dialogikontekstissa.

(Arno) Tämä esikoulutus perustuu siihen, että mallille näytetään tekstiä ja sitä pyydetään ennustamaan sitä seuraavaa sanaa. Tuntuu melko suurelta harppaukselta, että yhtäkkiä tällainen malli lopulta sitten osaa todistaa uusia matemaattisia lauseita tai ohjelmoida vaikka uuden pelin käyttäjän kuvauksen perusteella. Miten siihen päästään näillä malleilla?

(Sampo) Joo, tästä uutuudesta voitaisiin keskustella aika pitkäänkin. Nämä mallithan ovat selvästi erinomaisia interpoloimaan sen datan sisällä, jonka ne ovat nähneet koulutuksen aikana. Eli jos ne ovat esimerkiksi lukeneet biolääketieteellisiä artikkeleita, niin ne tietävät jotain tästä. Jos ne ovat lukeneet tai nähneet runollisia tekstejä koulutusvaiheessa, niin ne pystyvät tuottamaan sen tyyppistä tekstiä. Ja tämän kautta ne pystyvät sitten kirjoittamaan runomitassa biolääketieteellisistä aiheista. Se, että kuinka paljon nämä mallit pystyvät ekstrapoloimaan eli menemään tämän koulutuksessa näkemänsä datan ulkopuolelle ja tekemään täysin uutta, se on vielä jokseenkin epäselvää ja siinä on aika paljon haasteita. Eli mitä vähemmän sen tyyppistä dataa mitä malli on koulutuksensa aikana nähnyt, niin kun mennään sellaisille alueille tätä jakaumaa, niin sitä todennäköisempää on, että malli tekee jotain virheitä tai hallusinoi. Ehkä kieli on tässä yksi hyvä esimerkki. Kaikki aika varhaisetkin kielimallit alkoivat nähdä ikään kuin vahingossa jonkun verran esimerkiksi suomen kieltä, vaikka tarkoitus olisikin ollut tehdä puhtaasti englanninkielinen malli. Koska mikään miljardien tai biljoonien sanojen aineisto ei voi olla puhtaasti yksikielinen. Mutta jos aikaisemmilta englannin kieleen keskittyviltä malleilta pyysi suomenkielistä tulosta, niin tuotokset olivat usein aika heikkoja.

(Eric) Ajatteletko, että nämä mallit eivät pysty ekstrapoloimaan? Jos ajatellaan esim. viime aikoina tulleita raportteja siitä, että ne pystyvät todistamaan jonkun uuden matemaattisen lauseen, jota ei ole kukaan ihminen ennen onnistunut todistamaan, niin onko siinäkin loppukädessä kyse interpoloinnista? Että siinä on samoja osaongelmia kuin jossain tunnetuissa?

(Sampo) En uskaltaisi sanoa uskonko vai en. Sanoisin, että tämä on jokseenkin avoin kysymys. Ja kyllä, ehdottomasti uskon, että se, että uusi ongelma voidaan jakaa tutun näköisiin osaongelmiin, on osa sitä vastausta, minkä takia nämä mallit pystyvät tähän.

(Arno) Menee ehkä vähän filosofiselle sektorille, mutta kuinka usein me ihmisetkään keksimme mitään oikeasti uutta?

(Eric) Niin.

(Arno) Minä en ainakaan ole ihan varma, että olenko itse koskaan keksinyt mitään oikeasti uutta. Tietysti tieteenkin tekeminen on usein sitä, että yhdistellään asioita uudella tavalla. Ja varmaan sama pätee taiteessakin. Että tavallaan uusia oivaltavia tapoja, mutta se on silti yhdistämistä. Ja interpolointiahan sekin sitten on.

(Eric) Niinpä. No jos nyt vähän jatketaan tällä. Vähän väistit noita aiempia filosofisia kysymyksiä, mutta haluan kuitenkin kysyä vielä. Meidän podcastimme nimi on Ja ihminen loi älyn. Olemmeko me mielestäsi luoneet älyä vai vaan tehokkaan tavan jäljitellä jotain älykkään kuuloista tekstirakennetta?

(Sampo) No tämä on juuri se kysymys, jota yritän yleensä väistää. Ehdottomasti ollaan luotu jotain älykkään kaltaista. On selvää, että nämä mallit tekevät sellaisia asioita, jotka edellyttäisivät älyä, jos ihminen tekisi ne. Se, että missä määrin nämä mallit ovat älykkäitä, ja erityisesti se, että missä määrin niiden mahdollinen älykkyys on ihmisälyn kaltaista, on vaikea kysymys, johon en oikein haluaisi ottaa kantaa.

(Arno) Minkä kaltaista ihmisäly on? En ole ihan varma.

(Eric) Niin, sitä minä mietin myös.

(Sampo) Tarvittaisiin jonkun muun alan tutkija kyllä vastaamaan tähän.

(Arno) Siinä ehkä idea tuleviin jaksoihin.

(Eric) Joo.

(Arno) Hyvä. Siirrytään puhumaan kehittämistänne kielimalleista suomen kielellä kuten FinBERT ja Poro. Aloitetaan siitä, mikä tekee suomen kielen käsittelystä haastavaa kielimalleilla? Data? Vai onko suomen kielessä jotain muuta, joka tekee asioista vaikeaa?

(Sampo) Joo, me uskoimme pitkään, emme ainoastaan me TurkuNLP:ssä, vaan kielentutkijat laajemmin, että suomen kieli on erityisen haastavaa. Tämä osoittautui todeksi erityisesti sääntöpohjaisille malleille, joille suomenkielen morfologian, siis kielen taivutuksen, sanojen taivutuksen haasteet olivat yksi haaste, joka motivoi sitä, että Suomessa kehitettiin tiettyjä sääntöpohjaisia menetelmiä. Erityisesti vastaamaan sellaisiin haasteisiin, joita esiintyi suomen kielessä. Verraten siis englannin kielen taivutussääntöihin, jotka ovat triviaaleja. Ne kirjoittaa kuka tahansa, kun hetken aikaa asiaa ajattelee, niin iltapäivässä ylös. Suomen kielelle meni aika paljon pidempään, että päästiin siihen sääntöpohjaisilla menetelmillä. Meillä TurkuNLP:ssähän on ollut pitkäjänteisesti koneoppimispohjainen lähestymistapa kielen käsittelyyn. Siellä niin pitkään kuin vielä touhuttiin eksplisiittisten piirteiden parissa, että rikottiin sitä tekstiä

ihmisten kirjoittamien sääntöjen mukaisesti piirteisiin, joita koneoppimismenetelmä sitten pystyi käyttämään, niin vaikutti samalta, että täytyy olla suomen kielen erityisymmärrystä ja erityisosaamista, jotta pystyy kehittämään menetelmiä, jotka toimivat suomen kielellä.

Mutta tämä on muuttunut aika paljon nyt syväoppimisen ja piirreoppimisen kautta. Eli tällä hetkellä oikeastaan alkaen FinBERTistä ja osittain jo aikaisemmin TurkuNLP:ssä kehitettyjen suomen kielen neuroverkkopohjaisten jäsentimien yhteydessä me emme oikeastaan enää tee mitään kielispesifistä. Eli kaikki tulee datan kautta. Se, että meidän menetelmämme, FinBERT, FinGPT jne., jossain mielessä ymmärtävät tai tuottavat suomen kieltä, niin on kaikki siinä datassa. Menetelmällisesti nämä ovat täysin samoja. Totta kai me niin kuin teemme sellaisen tokenisoijan, joka on erikoistunut suomen kielelle, että se pystyy tehokkaasti esittämään suomen kielen palasia. Mutta sekin on menetelmällisesti ihan samannäköinen. Ihan samalla tavalla opitaan tämä tekstin merkkijonon jakaminen tokeneiksi tai sanan kaltaisiksi kappaleiksi kuin opittaisiin englannin, japanin tai minkä tahansa muunkielisen tekstin jakaminen. Eli nykyään pelataan datalla.

(Eric) Mainitsit FinBERTin ja FinGPT:n. Puhutaan näistä lisää. Jos vähän taustoitin, niin silloin, kun transformer-mallit, jos oikein muistan, julkaistiin 2017, olin itse kouluttamassa niitä ensimmäisiä transformer-malleja joihinkin omiin käyttötarkoituksiini ja muistan, että se oli todella erilainen kokemus. Ne hallusinoivat tosi paljon enemmän. Sieltä tuli paljon kielioppivirheitä. Mikä on ollut keskeinen ratkaisu tähän, mihin sinunkin tutkimuksesi läheisesti liittyy, on ollut nimenomaan tämä esikoulutus-idea. Eli että näille malleille syötetään paljon jotain sellaista aineistoa, joka ei liity välttämättä siihen tehtävään, johon se halutaan kouluttaa. Mutta että ne rakentavat jonkun mallin maailmasta ja ymmärryksen siitä kielen rakenteesta. Ja BERT-malleissa, jotka tulivat 2018, se perustui siihen, että ei ennustetakaan seuraavaa sanaa, vaan keskeltä jotain tekstiä piilotetaan joitain sanoja ja sitten pyydetään mallia ennustamaan ne. Tämä oli tosi iso läpimurto. Sillä oli tosi paljon käyttösovelluksia tällaisiin erilaisiin luokittelutehtäviin. Ei niinkään ehkä uuden tekstin generointiin. No, teidän FinBERT-mallinne julkaistiin sitten jonkin aikaa sen jälkeen. Teillä on artikkeli nimeltään Multilingual is not enough. BERTistä oli jo olemassa monikielinen versio, mutta te pystyitte osoittamaan, että teidän mallinne pystyy pärjäämään selkeästi paremmin suomenkielissä taskeissa. Mikä tässä oli se oivallus? Tai miten te saitte kehitettyä paremman mallin kuin nämä Googlen kehittämät monikieliset mallit?

(Sampo) Meillä oli aika pitkälti sama resepti kuin Googlellakin. Eli taas kerran datan kautta. Siinä vaiheessa, kun transformerit alkoivat lyödä läpi, ja oli erittäin hyviä transformer-pohjaisia BERT- ja GPT-malleja tarjolla monille suurille kielille, niin oli pitkään sellainen aukko, jossa nämä isoimmat kansainväliset toimijat eivät paljoa tuntuneet välittävän pienemmistä kielistä, kuten suomesta. Vielä alkuperäisen GPT-nelosen aikoihin oli selkeästi havaittavissa, että ei tässä nyt ole ainakaan ketään sellaista ollut mukana, joka erityisesti välittäisi suomen kielestä. Ja me välitimme suomen kielestä. Joten siis me teimme aika paljon datatyötä. Olimme tehneet jo aikaisemmin. Tässä oli näiden ensimmäisten mallien kehityksessä suurena apuna se, että TurkuNLP:n professorit Filip Ginter ja Veronika Laippala olivat jo pitkään pitäneet

käynnissä tällaista projektia, jossa he keräsivät internetistä erityisesti laadukasta suomenkielistä aineistoa. Ja sitten se oli osa sitä ensimmäistä koulutusaineistoa, jonka me koostimme sitten näiden FinBERT- ja FinGPT-mallien koulutukseen. Ja tietysti tässä ehkä se erikoisosaaminen, jota ei välttämättä monelta suurelta kansainväliseltä toimijalta löytynyt aktiivisesti, oli että me ymmärsimme suomen kieltä. Me pystyimme lukemaan suomenkielistä tekstiä ja sanomaan, että okei, tämä todennäköisesti on hyvää koulutusmateriaalia tai tämä on jotain huonon kielenkäännöslaitteen läpi mennyttä roskaa, jota me emme halua meidän koulutusdataamme.

(Eric) Oliko siinä enemmän kyse juuri tästä datan laadusta vai määrästä?

(Sampo) Uskoisin, että molemmista. En muista kylmiltään, kuinka paljon suomen kieltä olisi ollut multilingual BERTin, tämän Googlen alkuperäisen multilingual BERT -mallin koulutusdatassa, mutta olisin yllättynyt, jos olisi enemmän kuin muutamia promilleja. Meillä se oli 100 prosenttia. Meillä oli alussa vielä FinGPT:n aikoihin aika iso painotus datan laatuun. Me kehitimme ja sovelsimme paljon sääntöpohjaisia ja koneoppimispohjaisia menetelmiä siihen, että sieltä suodatettiin pois sellaista tekstiä, jota me pidimme epätoivottavana. Tämä ehkä pohjasi sitten vähän tällaiseen vanhanaikaisempaan ajatteluun siitä, mitä hyvän koneoppimismenetelmän koulutukseen tarvitaan. Että muutamakin huono esimerkki voi harhaanjohtaa tällaisia tietynlaisia edellisten sukupolvien menetelmiä. Nytemmin ehkä sitten esimerkiksi nykyisen OpenEuroLLM-projektin yhteydessä ollaan siirrytty vähän enemmän tällaiseen ymmärrykseen, että jossain määrin määrä voi korvata laatua. Erityisesti kun puhutaan pienistä kielistä kuten Suomi, joissa haasteena on se, että saadaanko ylipäättänsäkään mitään epätriviaalia määrää sitä kieltä siihen koulutusdataan. Eli tässä kun tehdään pohjamalleja, niin nyt vähemmän aggressiivisesti suodatetaan, jotta saadaan vain enemmän ja laajemmin erityyppistä suomen kieltä. Ja tietysti kaikkia muitakin kieliä siihen pohjamallin alkuperäiseen esikoulutusdataan.

(Eric) Jos käyttäjänä käytän jotain kielimallia, niin en välttämättä kirjoita aina kovin kielioopin mukaista tekstiä. Että siinä on tärkeääkin, että se malli osaa myös ymmärtää sitä.

(Sampo) Kyllä, ehdottomasti juuri tämä. Meillä oli alussa sellainen ajatus. Siis minun tutkimukseni tausta, jos mennään vähän kauemmaksi, on juuri biolääketieteen tieteellisen tekstin analyysissä. Tämähän on hyvin editoitua, huolellisesti kirjoitettua, monta kertaa katsottu läpi, ennen kuin mitään tällaista julkaistaan. Sieltä sitten ehkä osittain tuli tällainen ajatus, että me haluamme meidän malleillemme antaa vain ns. huippulaatuista tekstiä. Ja tosiaan, koska nämä mallit, niin kuin aikaisemmin juteltiin, toimivat parhaiten, jos ollaan sellaisessa alueessa sitä kielijakaumaa, jota ne ovat nähneet joko esikoulutuksessa tai myöhemmin, niin on tärkeää, että se koulutusdata sisältää myös rikkonaista kieltä, puhekieltä jne.

(Arno) Tosiaan tämä FinBERT oli tavallaan suomenkielinen projekti. Ja tavallaan siellä oli juuri se, että ei riitä, että on tällainen multilingual, monikielinen. Mutta sitten teillä myöhemmin on ehkä suunta muuttunut. Että se onkin tavallaan ehkä etu, että on monenkielistä materiaalia. Poro esimerkiksi on ilmeisesti hyötynyt tästä monikielisestä

materiaalista myös suomen kielen suhteen. Miten se kielten välinen dynamiikka tässä treenaamisessa toimii? Pystyykö malli jotenkin lainaamaan sitten siinä treenauksen aikana muista kielistä ja jotenkin tavallaan yleistämään asioita helpommin?

(Sampo) Joo, näin toivotaan. Jos voin hetkeksi palata tähän suunnanmuutokseen, niin siihen on hyvin yksinkertainen syy. Suomen kieli loppui kesken. Eli vielä kun tehtiin tätä alkuperäistä FinBERTiä, niin tällaisille pienemmille BERT-malleille oli silloin ennen vuotta 2020 vielä mahdollista kouluttaa tällainen malli nolasta muutamalla miljardilla sanalla. Ja muutama miljardi sanaa suomea oli vielä kohtuullisen helppoa laittaa kasaan siinä vaiheessa, kun tehtiin FinGPT:tä. Joitain vuosia myöhemmin tarvittiin jo satakertainen aineisto, 300 miljardia. 300 miljardilla sanalla koulutettiin. Ja siinä alettiin jo toistaa samaa koulutusmateriaalia monta kertaa, koska meillä ei ollut niin paljon. Meillä oli muutamia kymmeniä miljardeja sanoja suomea, vaikka pistettiin kasaan enemmän tai vähemmän kaikki mitä löydettiin. Ja sitten kun alettiin tehdä Poroa, niin oli selvää, että kunnianhimo oli noussut jokseenkin. Lumi oli käytettävissä, ja meillä oli tavoitteena kouluttaa biljoonalla sanalla tämä malli. Siihen ei enää suomen kieli venynyt. Sitä aineistoa, jonka saimme kasaan, olisimme joutuneet toistamaan niin monta kertaa, että oli hyviä syitä uskoa, että tämä malli olisi alkanut kärsiä siitä. Se olisi alkanut vain muistaa ulkoa sitä tekstiä, jota se on nähnyt uudestaan ja uudestaan. Sen sijaan, että se pystyisi edes jossain määrin interpoloimaan tai ehkä jopa ekstrapoloimaan.

(Eric) Mainitsit Lumin. Voitko selittää sellaiselle, joka ei tiedä, mistä on kyse, mikä tämä sinun Lumisi on, ja mikä sen merkitys on Suomelle?

(Sampo) Kunpa olisikin minun Lumini. Lumi on Euroopan, EU:n ja monen Euroopan maan konsortion supertietokone. Se oli vielä reilu vuosi takaperin yksi Euroopan suurimmista ja erityisesti EuroHPC:n, siis Euroopassa suurteholaskentaa hallinnoivan EU:n organisaation, supertietokoneista suurin. Ja se Lumi on ollut tämä laskentateho melkein kaikkien näiden meidän suurempien malliemme takana. Eli tämä FinBERT tehtiin vielä edellisen sukupolven pienemmällä CSC:n, siis suomalaisen tieteellisen laskennan keskuksen kansallisella supertietokoneella. Ja siinä vaiheessa, ennen kuin Lumi tuli Suomeen, niin meillä olisi ollut kunnianhimoa yrittää tehdä generatiivisia malleja, kuten GPT, mutta Suomesta sitä laskentaa ei löytynyt. Ja suuri osa siitä, mikä mahdollisti tämän alkuperäisen FinGPT-perheen oli se, että me saimme pilottiprojekti Lumille silloin, kun se järjestelmä Suomessa ensimmäistä kertaa käynnistettiin. Silloin pystyttiin ensimmäistä kertaa yrittää tehdä jotain näin suurta. Koska se laskentateho, jota tarvittiin, oli satoja tuhansia ellei miljoonia GPU-tunteja. Sellaista nyt vasta Lumin tullessa pystyttiin ensimmäistä kertaa Suomessa yrittämään.

(Eric) Voidaan tiivistetysti sanoa, että sillä on ollut tosi merkittävä vaikutus näiden suomenkielisten mallien kehitykselle. Monesti puhutaan siitä, miten laskentakapasiteetti on pullonkaula, ja miksi akatemiassa on hankala tehdä tällaista johtavien mallin kehitystä, koska ei ole sitä kapasiteettia. Mutta Suomessa on ollut poikkeuksellisen hyvät resurssit tähän.

(Sampo) Kyllä kyllä. Aivan ehdottomasti. Mikään tästä meidän myöhemmästä työstämme ei olisi ollut mahdollista, ellei meillä olisi ollut tätä tilaisuutta sen kautta, että Suomessa on ollut yksi Euroopan suurimmista supertietokoneista, ja suomalaisella tutkimuksella on ollut erittäin suuri osa siitä kokonaiskapasiteetista.

(Eric) Puhuit tästä koulutusdatan loppumisesta. Se voi ehkä kuulostaa yllättävältä, jos ajatellaan, että eikö koko ajan synny kuitenkin uutta tekstiä, uusia uutisartikkeleita ja ihmiset puhuvat sosiaalisessa mediassa. Voisiko koittaa vähän vielä avata yleisesti, että miksi tämä on ongelma? Ja miten on mahdollista, että se data pääsee loppumaan?

(Arno) Niin, miten niin loppuu kesken? Kyllähän tässä koko ajan puhutaan suomea nytkin.

(Sampo) Joo, ehkä aluksi voisin sanoa alta pois, että suuret kansainväliset toimijat ehkä käyttävät uutisia ja sosiaalisen median aineistoja, jos heillä on jotain kautta sellaiseen pääsy. Me koulutamme ainoastaan julkisesti saatavilla aineistoilla, emmekä siis käytä esimerkiksi sosiaalista mediaa enää näissä uusimmissa malleissa. Mutta haasteena on se, että tuotetun tekstin määrä kasvaa pääosin lineaarisesti. Eli ihmiset tuottavat tekstiä suurin piirtein samaa tahtia kuin ovat tuottaneet viime vuonna tai sitä edeltävänä. Ja mallien koulutukseen käytetty datan määrä on kasvanut eksponentiaalisesti. Eli näitä meidän mallejamme esimerkkinä käyttäen, niin FinBERTissä taisi olla 3 miljardia, FinGPT:ssä 30 miljardia. Sitten nyt aletaan puhua sadoista tai miljardeista tai biljoonista sanoista, mitä käytetään tällä hetkellä näiden OpenEuroLLM-mallien kouluttamiseen. Tällaisella eksponentiaalisella käyrällä ei yksinkertaisesti voida pysyä kovin pitkään, ennen kuin hyvinkin suuret dataresurssit loppuvat. Nyt aletaan olla siinä pisteessä, että avoimesti saatavilla olevien internetsivujen sisältämä data on jo rajoittava tekijä sille, mitä pystytään tekemään.

(Arno) Mitä sitten, jos saisit itse valita unelma-aineiston? Kirjastot pullistelevat kirjoja. Varmaan Ylellä, Hesarilla ja Aamulehdellä on isot arkistot suomenkielistä hyvää dataa. Onko tämä kaikki käytettävissä? Vai miten tämä toimii käytännössä? Puhuit avoimesti saatavilla olevista. Mitä se käsittää käytännössä?

(Sampo) Niin, meidän perspektiivistämme valitettavasti, oikeudenhaltijoiden perspektiivistä ehkä sitten toivottavasti, niin esimerkiksi monet uutisorganisaatiothan ovat viime vuosina kommunikoineet kohtuullisen selvästi, että he eivät toivo, että heidän materiaaliaan käytetään kielimallien kouluttamiseen. Riippumatta siitä, onko siitä ollut osia avoimesti internetissä saatavilla. Tästä johtuen me emme enää nykyään käytä tällaista dataa. Sama meillä oli FinGPT:n aikoihin. Oli tällainen toive, että muistiorganisaatiot, kirjastot, kansallisarkistot ja muut, voisivat tukea tällaisten kansallisten mallien kehitystyötä. Ja meillä oli jonkin aikaa yhteistyötä Kansalliskirjaston kanssa erityisesti näiden ensimmäisten FinGPT-mallien kouluttamisessa. Tämä on sittemmin nähty vähemmän toivottavaksi. Tällä hetkellä tehdään puhtaasti sellaisen aineiston pohjalta, joka on kenelle tahansa internetistä vapaasti ladattavissa. Aina kun mahdollista, niin tietysti myös sellaisen kautta, jolla on avoimet lisenssit. Tämä on meille jokseenkin rajoittava tekijä meidän työssämme. Jos katsotaan, miten suuret kansainväliset toimijat keräävät oman aineistonsa, niin säännöt vaikuttavat olevan

vähän toiset. Esimerkiksi nyt tämä Anthropicin kirjojenpolttamisoperaatio on jotain, johon emme olisi ryhtymässä. Jos kaikki tällainen kansallinen materiaali, uutiset, kirjastot jne. olisi saatavilla, varmasti pystyttäisiin parempaan. Mutta no, se ei tässä nykytilanteessa ole meille enää mahdollista. Eli pelataan lähinnä sen varassa, mitä julkisessa internetissä on saatavilla.

(Arno) Ehkä peräänkuuluttaisin kansallista missiota myös suomen kielen tukemiseksi ja turvaamiseksi tässä näin. Tuntuu vähän yksisilmäiseltä estää se, varsinkin esikoulutuksen suhteen. Eri asia on sitten tuotteistaa ja myydä tavallaan. Mutta se esikoulutus on kuitenkin sitä, että ymmärretään kielen rakennetta, rakennetaan malleja, meillä on hyvä suomenkielinen pohja. Eikö se olisi kansallisestikin tosi tärkeä asia?

(Sampo) Tämä on minun näkemykseni. Kyllä minä ymmärrän molempia perspektiivejä. Ja totta kai me kunnioitamme näiden tekstien kirjoittajien toiveita silloinkin, kun me ehkä toivoisimme, että heidän näkemyksensä olisi toinen. Tämähän on nähty, siis Islannissa ja Virossa muistaakseni on, en tiedä, on varmaan muitakin esimerkkejä, mutta on tuotettu tällaisia kansallisia aineistokokoelmia, joita on sitten tarjottu kielimallien koulutukseen. Erityisesti sen takia, että nämä kielet säilyisivät vahvempina näissä malleissa.

(Eric) Niin, mietin, että osaatko konkretisoida mitä ne edut voisivat olla, jos nämä olisivat käytettävissä? Että mihin näitä malleja esim. voitaisiin soveltaa sitten, mikä olisi Suomelle arvokasta ja tärkeää?

(Sampo) Sovellukset olisivat varmaan samantyyppisiä. Eli kyse ei ole niinkään kategoriaerosta, vaan siitä, että karkeasti ottaen mitä enemmän, mitä monipuolisempaa ja mitä laadukkaampaa tekstiä kullekin kielelle tässä mallin koulutusaineistossa on, niin sitä paremmin se malli pystyy sitä käsittelemään.

(Arno) Tässä vähän ristiriitaisia viestejä tulee. Samalla halutaan terveydenhuoltoa ja muuta digitalisoida ja tuoda tekoälyä sinne, mutta sitten se varmaan pitäisi tapahtua suomen kielellä tai kotimaisilla kielillä. Sitten taas toisaalta mikään ei tue kotimaisten kielten mallien koulutuksessa datan saatavuutta ja muuta. Niin ehkä tässä olisi poliittisen keskustelun paikka, että haluammeko me turvata myös tulevaisuudessa suomen kieltä. Haasteena on osittain myös se, että koska suomenkielistä materiaalia, riippumatta siitä, kerätäänkö sitä kirjastoista vai pelkästään netistä, niin on kaiken kaikkiaan niin vähän, että enää nykyään ei ole mielekästä yrittääkään tehdä yksikielisiä malleja tai ei ainakaan suuria yksikielisiä malleja. Tämä FinGPT:hän oli tosiaan meidän viimeinen puhtaasti suomenkielinen. Sen jälkeen, kun haluttiin tehdä suurempaa, niin Poro tehtiin jo yhdistelmällä suomi-englanti ja jonkin verran koodia sekaan. Ja nykyään tehdään jo OpenEuroLLM:ssä malleja 30-40 eri kielelle. Ja siinä vaiheessa tämäntyyppiset erityistapaussopimukset, joissa esimerkiksi päästäisiin suomenkielisiin kansallisiin arkistoihin potentiaalisesti jollain tavalla hyötymään heidän materiaaleistaan, niin tulee jo paljon vaikeammaksi, koska tämä pitäisi tehdä 30-40:ssä eri maassa. Eli ehkä, no, poliittisiin kysymyksiin yritän olla ottamatta kantaa, mutta jos jotain tällaista tapahtuisi, niin sen pitäisi tapahtua, sanotaan, ainakin EU:n tasolla, ei niinkään kansallisella.

(Arno) Kuulostaa järkevältä. Olet mukana sellaisessa tutkimusartikkelissa kuin Scaling Data-Constrained Language Models. Sen mukaan datan käyttäminen useampaan kertaan voi olla järkevää tai hyödyllistä. Mitä riskejä tähän liittyy ja mitä hyötyjä?

(Sampo) Joo, tämä on ollut meille tosi tärkeä kysymys. Erityisesti, koska toimitaan suomen kielen parissa, ja datan määrä on ollut jatkuvasti rajoite. Eli edellisen sukupolven menetelmien parissa toimiessa oli aina selvää, varsinkin kun tehtiin ohjattua oppimista, jossa koulutettiin paljon pienemmällä määrällä käsin annotoitua dataa, ja yksittäisen datapisteen arvo oli paljon suurempi, että totta kai halutaan tehdä muutaman epokki. Siis käydä data läpi muutamaan kertaan. Koska neuroverkkopohjaiset menetelmät oppivat lopulta yllättävän vähän joka kerta, kun ne näkevät uuden datapisteen. Sen takia silloin, kun data on vähäistä tai arvokasta, niin on selvästi järkevää käydä jokainen datapiste läpi useamman kerran tämän opetusprosessin aikana. Ja tämä, kuinka monta kertaa opetusdataa toistetaan sen koulutusprosessin aikana, niin on tyypillisesti ollut yksi ikään kuin koulutuksen hyperparametri eli sellainen mitä itse tutkija sitten säätää. Kun yritetään optimoida tätä koulutusprosessia. Silloin kun oltiin hetken aikaa datarikkaita, kun alettiin toimia internet-skaalan aineistojen parissa, niin tämä ajatus vähän unohtui aika monelta toimijalta. Alettiin katsoa, että datan määrä on rajaton, joten otetaan aina uutta dataa sen sijaan, että toistettaisiin sitä mitä ollaan näytetty menetelmälle jo aikaisemmin. Koska uuden datapisteen informaatio sisältö on kuitenkin aina vähän enemmän kuin sellaisen, jonka se malli on jo nähnyt. Koska se on oppinut jo ainakin jonkun osan siitä informaatiosta, jota se on nähnyt aikaisemmin. Osittain tämän motivoimana teimme tämän artikkelin, jossa tutkitaan sitä, että silloin, kun data alkaa taas olla rajoite, niin kuinka monta kertaa voidaan näyttää samaa datapistettä tällaiselle suurelle generatiiviselle neuroverkkomallille ennen kuin siirrytään siitä, että malli oppii jotain uutta siitä datasta, jonka se on jo kertaalleen nähnyt, siihen, että malli alkaa vain muistaa ulkoa sen datan. Koska siltä pyydetään vaan toistuvasti, että ennusta nyt juuri tämä seuraava sana juuri tässä kontekstissa. Ja tästä työstä nyt on tullut sellaisia karkeita ohjenuoria, joita me vielä tänään noudatamme. Että sanotaan muutamia kertoja, ehkä 4-6 kertaa voidaan käydä koulutusdata läpi. Jos mennään kymmeniin kertoihin, niin siinä vaiheessa mallin suorituskyky alkaa jo selvästi kärsiä ja se, että saako se datasta enää mitään uutta irti, niin on aika haasteellista. Kun ollaan eksponentiaalisella käyrällä tai yritetään pysyä eksponentiaalisen kasvun tahdissa siinä, että kuinka paljon dataa näytetään näille malleille, niin tämä, että voidaan 4-6 kertaa toistaa kaikki data, voi antaa meille melkein yhden askeleen lisää. Mutta tietenkin tämä on kertakäyttöinen hyöty. Mutta tämä on nyt ollut meillä ohjenuorana sitten myöhemmässä työssä. Ja tosiaan FinGPT-, Poro- ja nykyisissä OpenEuroLLM-malleissa pienempien kielten, kuten suomen, koulutusdataa toistetaan juuri tämän tyyppisesti.

(Eric) Ehkä liittyen datan uudelleenkäyttöön, jos en väärin muista, niin olen lukenut jonkun tilaston, että jo yli puolet uudesta internet-materiaalista olisi jo tekoälyn tuottamaa. Eikö tämä ole ongelma, että sitten syötetään sitä tekoälyn omaa outputtia sille itselleen?

(Arno) Vai ratkaisu, koska dataa on enemmän? Emmekö me vain generoi enemmän dataa?

(Sampo) Voidaan generoida enemmän dataa. Tähän liittyy tosiaan suuria mahdollisuuksia ja suuria riskejä. Joitain vuosia takaperin oli aika vaikutusvaltainen artikkeli, joka tutki erityisesti tätä mallin romahtamista tällaisessa tapauksessa, jossa aina seuraavan sukupolven malli koulutetaan edellisen sukupolven mallin generoimalla datalla. Tässä työssä osoitettiin, että ei tarvitse kovin montaa kertaa iteroida tätä, ennen kuin tämä mallin tuottama tai mallintama todennäköisyysjakauma ei enää muistuta millään hyödyllisellä tavalla sitä alkuperäistä dataa. Osittain tämän takia tähän on liittynyt aika paljon pelkoja, että jos koulutusdataan pääsee sekoittumaan mallien tuottamaa dataa, niin missä määrin se voi mahdollisesti myrkyttää näitä malleja? Tällä hetkellä ymmärrykseni on, että niin pitkään kuin koulutusdatassa on kuitenkin pääosa ihmisten tuottamaa, jossa siis on sitä entropiaa, jota tarvitaan, että mallit eivät täysin romahda, niin, se, että data sisältää jonkin verran muiden mallien generoimaa tekstiä, ei välttämättä ole suurikaan katastrofi. Sitten tietysti tällä positiivisella puolella, jos sen sijaan, että koulutusdataan vahingossa lipsahtaa mukaan muiden mallien generoimaa tekstiä, niin voidaan huolellisesti valita sitä, miten generoidaan koulutusdataa, niin tässä on suuriakin mahdollisuuksia. Ja kyllä tällainen synteettinen data nähdään nykyään yhtenä potentiaalisena vastauksena siihen, että mitä tehdään, kun datamuuri sitten lopulta tulee kunnolla vastaan. Eli sitten kun ihmisten tuottama teksti ei enää riitä, niin sitten toivotaan, että löydetään sellaisia menetelmiä, joilla voidaan tuottaa riittävän monipuolista, riittävän laadukasta synteettistä dataa, jotta sen avulla voitaisiin ehkä ottaa tämä seuraava kertaluokka-askel, kun tarvitaan taas 10 kertaa enemmän.

(Arno) Mutta onko nyt jotenkin varmasti tallessa jotain luomudataa, jossa sitä ei ole? Jos sanotaan vaikka, että se, että sinne sekoittuu sitä tekoälymallien luomaa dataa on vähän niin kuin joku kävisi kusemassa sinne sinun kaivoosi. Niin onko jossain sitä puhdasta vettä tallessa? Sitä aidosti ihmisten tuottamaa tekstiä?

(Sampo) Joo, internet-aineistojen tapauksessa tähän on onneksi aika yksinkertainen vastaus. Eli jos lakataan vuodesta 2023 eteenpäin keräämästä, siis otetaan vain ne internet-crawlit, jotka tehtiin ennen vuotta 2023, niin ollaan kohtuullisen varmoja, että se on hyvinkin suurilta osin ihmisten tuottamaa, mitä nyt seassa on huonoja konekäännöksiä.

(Eric) Ehkä tästä tulee mieleen yksi ongelma myös. Mietitään vaikka opiskelijoiden vastausten tarkistamista. Voidaanko luotettavasti tunnistaa, mikä on tekoälyn tuottamaa tekstiä ja mikä ei?

(Sampo) Tämä ei ole minun erityisosaamisalaani, mutta ymmärrykseni on, että jos on tarpeeksi suuri aineisto ja ymmärrys siitä, mitkä mallit olisivat sen voineet potentiaalisesti tuottaa, niin voidaan. Mutta yleisemmin kysymykseen siitä, onko tämä teksti, varsinkin jos on verraten lyhyt tekstinpätkä, jonkun mahdollisesti tuntemattoman mallin tuottamaa, niin tähän kysymykseen ei voida luotettavasti vastata.

(Eric) No, jos mietitään vielä nimenomaan suomenkielisen datan loppumista, niin yksi ratkaisu, jota tekin olette tutkineet, on konekääntäminen. Eikö voida vaan ottaa kaikkea sitä englanninkielistä dataa, ajaa se jonkun Google translate -tyyppisen ratkaisun läpi ja sitten kouluttaa se malli sillä? Ratkaiseeko tämä meidän ongelmamme?

(Sampo) Tämä auttaa, joo. Me olemme tosiaan tätä tällaisessa OpenEuroLLM:n ja muutaman muun toimijan yhteistyöprojektissa kuin MultiSynt, tutkineet. Saimme sille aika mukavasti laskentaa EuroHPC:ltä. Olemme nyt konekääntäneet muutamallakin eri menetelmällä suomen kielelle ja kymmenille muille Euroopan kielille 100 miljardia sanaa kohtuullisen korkeatasoista englanninkielistä tekstiä. Eli meillä on nykyään enemmän konekäännettyä suomea kuin ihmisten kirjoittamaa suomea, ja tämä laatuksymys on vaivannut meitä paljon. Me olemme tätä lähestyneet monesta perspektiivistä. Yksi ensimmäisistä havainnoista oli, että jos asetetaan kokonaan sivuun ihmisten tuottama data ja koulutetaan mallia puhtaasti tällä käännetyllä datalla, joka on siis käännetty erittäin pienestä tarkasti filitroidystä osasta englanninkielistä internet-aineistoa, niin meidän mallimme pärjäävät tällaisissa, sanotaanko vähän niin kuin naiiveissa monivalintaevaluaatioissa paremmin. Eli jos hyvinkin naivistisesti evaluoidaan näitä malleja, niin näyttää siltä, että ei tarvita suomen kielen puhujia ollenkaan. Tarvitaan laadukasta englanninkielistä aineistoa ja hyvä konekääntämismenetelmä. Mutta samalla, kun me vähän kokeilemme itse näitä generatiivisia malleja, jotka on koulutettu ainoastaan käännetyllä datalla, niin kyllä suomen kieltä puhuvalle ihmiselle on välittömästi selvää, että tämä ei ole ihan sama juttu. Tämän eron tunnistaminen automaattisesti on osoittautunut yllättävän hankalaksi. Tämä on siis meillä edelleenkin aktiivinen tutkimusaihe, ja tätä on välillä vähän vaikea selittää ihmisille, joiden fokus on ehkä enemmän isoissa kielissä. Että okei, nämä evaluaatiothan näyttävät siltä, että on parempi vaan käyttää tätä käännettyä dataa. Että ei meidän tarvitse ollenkaan keskittyä haalimaan tätä vaikeasti löydettävissä olevaa pientä määrää natiividataa. Me voimme vaan kääntää. Mutta uskon vahvasti, että tarvitaan edelleen suomen kieltä puhuvien ihmisten kirjoittamaa dataa. Tällä hetkellä me varovaisesti sekoitamme noin 30 prosenttia tällaista käännösdataa. Tähän numeroon päätyminen on ollut aika suuren väännön takana. Koska edelleenkin jos me vaan katsomme evaluaatiotuloksia, niin näyttäisi siltä, että oikea vastaus voisi olla 100 prosenttia käännösdataa.

(Eric) Tuo on jännä kysymys mielestäni senkin takia, että kielihän muuttuu koko ajan. Itse tulee varmaan käytettyä paljon anglismeja. Jos kuuntelen tytärtäni, niin hän käyttää ehkä sitäkin enemmän vielä. Ja nyt on ollut tätä myös, että kielimalleilla on jotkut tyypilliset tavat ilmaista joitain asioita, ja nekin vaikuttavat siihen jo, miten ihmiset ovat alkaneet kirjoittaa ja em dashia ja muuta käyttää omassa tekstissään. Joo, tosi mielenkiintoista.

(Arno) Puhuit siitä, että jotenkin siinä tekstissä tavallaan natiivipuhuja huomaa kuitenkin, että jokin ei oikein ole kohdallaan, helposti, jos siellä on liikaa tavallaan englannista käännettyä materiaalia. Mutta mikä siitä suomen kielen sielusta alkaa kadota ensin? Onko se sanojen taivutus vai sanajärjestys? Puhekieliset ilmaukset, idiomattiset ilmaukset, murre sanat vai kulttuurilliset? Mikä sieltä alkaa hävitä ensin?

(Sampo) Poimisin tästä listasta ehkä kaksi. Idiomaattiset ilmaukset ja kulttuurin ymmärtäminen. Eli nykyään konekääntämismenetelmät alkavat olla sen verran hyviä, että jos osaa valita kohtuullisen hyvän käännösmenetelmän, niin kyllä taivutukset ja kieliooppi pääosin menevät oikein. Sitten nämä erot alkavat näkyä vähän erilaisissa asioissa. Ja juuri tällaiset idiomattiset ilmaukset ovat yksi niistä, jota näkyy. Hauska yksityiskohta, joka näkyy selvästi, jos on koulutettu malli erityisesti konekäännetyllä materiaalilla, niin ihmisten nimet eivät ole suomalaisia nimiä. Paikat, joista se malli tykkää puhua, eivät ole Suomessa. Ja kun puhutaan Suomesta, suomen kielestä, suomalaisuudesta tai jostain tällaisesta, niin mallin perspektiivi, jos tätä sanaa nyt voidaan käyttää, on että katsotaan ulkoapäin sisään. Koska se teksti ei ole sisältänyt ajatusta "me suomalaiset", vaan enemmänkin, että "tuolla Suomessa".

(Arno) Nämä nykymallit eivät tee ihan yhtä räikeitä virheitä. Muistan eräänkin yhdysvaltalaismallin, joka minulle väitti, että kakkamakkara on suomalaista perinneruokaa. Tässä varmaan menivät kakkamakkara ja mustamakkara sekaisin. Ne eivät ole ihan näin räikeitä ne virheet enää. Mutta kyllä sen huomaa, että ei välttämättä ole fokuksessa näillä isoilla yrityksillä se suomenkielinen tuottaminen. Jos miettii kuinka kyvykkäitä nämä mallit ovat, ne pystyvät todistamaan matemaattisia lauseita, koodaamaan asioita. Mutta kuitenkin tämä suomen kieli jotenkin takeltelee edelleen. Se on vähän huolestuttavaakin jollain tavalla.

(Sampo) Joo, valitettavasti sama ilmiö alkaa tulla mukaan myös meidän malleihimme nykyään, kun meidän painotuksemme ei voi enää olla 100 prosenttisesti suomenkielessä. Olen joutunut havaitsemaan, että nyt kun koulutetaan 30-40:lle kielelle malleja, niin niiden suomen kielen tuottaminen on välillä selvästi heikompaa kuin näiden ensimmäisten 100 prosenttisesti suomenkielisten FinGPT-mallien. Tämän tasapainon löytäminen on aika vaikeata.

(Arno) Siirrytään puhumaan vähän avoimuudesta ja eurooppalaisista tekoälyasioista. Mikä on tärkein konkreettinen asia, jonka Suomi menettää, jos parhaat suomenkieliset tekoälypalvelut perustuvat vain muutaman ulkomaalaisen yrityksen suljettuihin malleihin?

(Sampo) Sama vastaus Suomelle kuin oikeastaan koko Euroopallekin, eli jonkinlaisen itsenäisyyden. Eli se, että jos me jäämme riippuvaisiksi Suomen ja Euroopan ulkopuolisista toimijoista siinä, että meillä on käytettävissä tätä teknologiaa omalla kielellämme, niin se tarkoittaa, että milloin tahansa nämä toimijat voivat rajata meidän käyttöämme. Tai no, jos katsotaan taloudellisesta perspektiivistä, niin se tarkoittaa, että me maksamme jatkuvasti veroa oman kieleemme käyttämisestä näissä palveluissa jonnekin Suomen ja Euroopan ulkopuolelle. Tätähän on povattu jo pitkään, mutta nyt viime vuosina tai tänä vuonna enemmänkin on alkanut näkyä ihan konkreettisia esimerkkejä siitä, että amerikkalaiset ja kiinalaiset toimijat alkavat rajoittaa pääsyä huippumalleihin. Siitä, mitkä motivaatiot ovat milloinkin kyseessä, taloudelliset, poliittiset tai muut, niin se on jokseenkin toissijaista. Mutta jos me haluamme taata itsellemme sen, että meidän kieltämme käyttävät mallit ovat jatkossa takuulla meille saatavilla, niin sitten meillä täytyy olla kapasiteetti tehdä näitä malleja itse. Myös siinä tapauksessa, että me emme välttämättä pysty ihan samassa kisassa olemaan. Eli

vaikka me tulemme vähän perässä, niin uskon, että on erityisen tärkeää, että me emme täysin luovu tästä kisasta. Eli me säilytämme vielä Suomessa ja Euroopassa kyvyn tarvittaessa tuottaa myös omia mallejamme sen sijaan, että siirtyisimme täyteen riippuvuuteen amerikkalaisista tai kiinalaisista malleista.

(Eric) Kun puhutaan avoimista kielimalleista, niin sillä saatetaan tarkoittaa vähän eri asioita. Minkä asioiden mielestäsi pitäisi olla avoimia? Joskus viitataan pelkästään sen mallin painoihin tai sitten koulutusdataan, siihen koodiin, jolla se on koulutettu ja siihen itse koulutusmenetelmään.

(Sampo) Uskon, että kaikille kategorioille on paikkansa. Ehkä jakaisin tätä maailmaa karkeasti ottaen kolmeen. Eli on nämä täysin suljetut mallit, kuten GPT:t ja Claudet, joiden koulutusdatasta, koulutusmenetelmästä tai välttämättä edes mallin koosta tai arkkitehtuurista ei oikeastaan julkaista yhtään mitään. Ainoa tapa olla minkäänlaisessa vuorovaikutuksessa näiden mallien kanssa on jonkin apin kautta. Ja näen, että tälle on oma paikkansa. On ymmärrettävää, että jos jotkut toimijat sijoittavat aika merkittävästi tällaisten mallien kehittämiseen, niin sitten voi olla mielekästä, että he pitävät joitain näitä asioita salaisina. Sitten keskellä tässä avoimuusskaalassa voidaan nähdä nämä LLaMA:t ja kumppanit, joiden painot ovat saatavilla. Eli voidaan ainakin periaatteessa ottaa se itse malli ja ajaa sitä lokaalisti maksamatta niistä tokeneista kenellekään muuten kuin sitten sähköön ja laitteiston kautta. Valitettavasti tässä yhteydessä puhutaan joskus aika harhaanjohtavasti avoimista malleista, vaikka olisi ainoastaan mallin painot saatavilla ja joskus ehkä vielä aika rajoittavin lisenssiehdoin. Tässä LLaMAN tapauksessa mallin arkkitehtuuri ja joitain sen koulutuksen yksityiskohtia on tiedossa, mutta datasta ei tiedetä paljoa mitään. Eli moninaisista syistä, joihin voidaan ehkä mennä, niin Meta on katsonut parhaaksi olla kommentoimatta, että millä datalla nämä heidän ns. avoimen lähteen mallit on koulutettu. Sitten tässä kolmeen jaossa avoimimmassa päässä olisivat mallit, joita voisin nimittää täysin avoimiksi. Tämä tarkoittaa sitä, että mallin painot ovat vapaasti saatavilla ja kenen tahansa käytettävissä mihin tahansa tarkoitukseen ilman rajoittavia lisenssi- tai muita ehtoja. Mallin koulutusdata on tiedossa ja vähintäänkin saatavilla tarkkailtavaksi, ja mallin koulutusprosessin yksityiskohdat, koodi jne. ovat saatavilla myös avoimin ehdoin. Eli siis avointa lähdekoodia, avoin prosessi. Ja jos kaikki tämä on saatavilla, niin silloin periaatteessa kuka tahansa, jolla on tämä tarvittava laskentakapasiteetti käytössä, niin pystyisi toistamaan tämän prosessin ja luomaan saman tai hyvin samantyyppisen mallin näistä samoista lähteistä. OpenEuroLLM:ssä me nyt toimimme tässä viimeisessä eli täysin avoimessa kategoriassa. Tämä on ollut meillä kauttaaltaan kunnianhimona meidän kielimallien kehitystyössämme TurkuNLP:ssä. Mutta aikaisemmin me suomenkielisen materiaalin rajallisuuden takia osittain päädyimme tekemään sellaisia kompromisseja, joissa me hyväksyimme, että jokin osa meidän koulutusdatastamme oli esimerkiksi rajoittavien lisenssien takana. Joka sitten johti siihen, että muut eivät välttämättä olisi pystyneet tuottamaan samaa mallia, vaikka me omasta puolestamme tietysti teimme kaiken mitä pystyimme sen eteen, että tämä olisi mahdollista. Minä uskon, että on erittäin tärkeää erityisesti tällaisen kansallisen ja EU-tason digitaalisen suvereniteetin perspektiivistä, että se mitä me teemme, on täysin avointa.

(Arno) Tämä OpenEuroLLM-projekti, jossa olet mukana, tavoittelee avoimia eurooppalaisia malleja kaikille EU:n virallisille kielille. Onko se ensisijaisesti tieteellinen hanke, teollisuuspolitiikkaa, kielipolitiikkaa vai digitaalista turvallisuuspolitiikkaa? Mikä ajaa tätä hanketta?

(Sampo) Motivaatioita löytyy varmaan kaikilta näiltä. EU:n järjestelmän perspektiivistä tämä on infrastruktuurihanke. Ei niinkään siis tutkimusta, vaan tuotetaan artefakteja, erityisesti siis kielimalleja. Meidän päävastuamme tässä projektissa, vaikka totta kai me haluamme tehdä uutta tieteellistä tutkimusta, niin tämän projektin pääasiallinen tavoite on kehittää näitä avoimia kielimalleja sekä siinä sivussa avoimesti dokumentoitu prosessi, jonka avulla voidaan jatkossa kehittää Euroopassa samankaltaisia ja toivottavasti parempia malleja. Tietysti tässä toivotaan myös, että sivutuotteena syntyy enemmän kielimalliosaamista Eurooppaan. Eli koulutetaan seuraavan sukupolven osaajia sitten, jotta he voivat jatkaa tämäntyyppistä työtä.

(Arno) Missä vaiheessa projekti etenee nyt?

(Sampo) Me olemme tällä hetkellä alun perin kolmivuotiseksi speksatun projektin suurin piirtein puolivälissä. Me olemme julkaisseet tusinoittain malleja, jotka ovat pääosin aika pienehköjä tutkimusprototyyppejä. Me olemme tällä hetkellä noin 80 prosentissa meidän ensimmäisen ns. tuotantomallimme koulutusta. Se on pienehkö tai sanotaanko keskikokoinen kielimalli. 8 miljardia parametria. 10 biljoonaa sanaa koulutusmateriaalia. Tämä tulee todennäköisesti, uskaltaisinko sanoa, noin kuukauden sisään saataville. Siitä on jo checkpointteja. Eli siis tällaisia ei täysin loppuun koulutettuja painoja on saatavilla jo 80 prosenttiin saakka. Ja kaikki työ tehdään täysin avoimesti. Näitä ns. keskeneräisiä malleja tai keskeneräisiä mallin painoja laitetaan sitä mukaa saataville avoimesti, kun niitä tulee. Tämän vuoden lopussa meillä on tavoitteena julkaista ensimmäinen ns. suuri kielimalli, joka vastaa näihin meidän keskeisiin tavoitteisiimme. Se tulee olemaan noin 30 miljardia parametria, koulutettu 15 biljoonalla sanalla. Se suurin piirtein liikkuu samassa kategoriassa kuin LLaMA:t ja DeepSeekit. Sitä, kuinka kaukana me olemme GPT:stä tai Claudesta laskennassa tai datan määrässä, on mahdotonta sanoa. Todennäköisesti he sijoittavat vielä vähän enemmän.

(Arno) Minua kiinnostaa ihan myös käytäntö, tavallaan, jos raotat vähän tätä mystisyyden verhoa, niin nehan ovat tosi isoja koulutusajoja, joita tässä joudutaan tekemään. Minäkin tavallaan olen kouluttanut ties kuinka monta mallia vuosien varrella, mutta ei mitään tätä kokoluokkaa. Mitä käytännön juttuja siinä pitää huomioida? Kun sinulla on monta GPU:ta, niin kai se rautakin saattaa hajota matkan varrella, ja kaikenlaisia käytännön ongelmia tulee varmaan vastaan.

(Sampo) Kyllä. Tämä näiden mallien kouluttaminen on aika paljonkin käytännön ongelmien ratkaisemista. Olen joissain yhteyksissä sanonut, että aikaisemmin tein tutkimustyötä ja nykyään roolini näyttää lähinnä olevan kiroilla supertietokoneille. Kun puhutaan sadoista tai tuhansista GPU:ista yhdessä koulutusajossa, niin kaikki, mikä mahdollisesti voi mennä pieleen, menee aika hyvää tahtia pieleen ainakin yhdelle näistä laitteista. Eli tosiaan tuhansilla GPU:illa toimitaan ja miljoonia GPU-tunteja

poltetaan kokonaisuudessaan näiden suurimpien mallien kouluttamiseen. Ja EuroHPC:n laitteistosta ehkä pitää ensimmäisenä sanoa, että me olemme erittäin kiitollisia siitä, että tällainen mahdollisuus ylipääntänsäkään on. Mutta EuroHPC:n supertietokonelaitteistoa ei ole suunniteltu ensisijaisesti tekoälymallien kouluttamiseen. Sellaisia laitteita alkaa nyt tulla seuraavassa sukupolvessa saataville. Mutta ne, joilla me koulutamme, mukaan lukien Lumi, niin ovat järjestelmiä, joiden alkuperäisessä suunnittelussa ei ole ollut ensimmäisenä tavoitteena se kielimallien koulutus, että ne ovat erityisen tehokkaita kielimallien koulutukseen. Ja se näkyy esimerkiksi näiden GPU:iden välillä olevissa kommunikaatioyhteyksissä, tiedonsiirron tehokkuudessa ja monessa muussa. Aika paljon tästä käytännön tehtävästä on sitten näihin haasteisiin vastaamista. Koetetaan selvittää yleensä sitten supertietokone-eksperttien, siis rautaeksperttien, kanssa yhdessä, että mikä on mahdollista, ja missä ne rajat ovat. Ja miten me ehkä voisimme tehdä jotain vielä vähän isompaa.

(Eric) Minua kiinnostaa tämän mittaluokan projektin toteuttaminen tai projektin koordinointi käytännössä. Mainitsit, että ihmisillä voi olla erilaisia intressejä, miksi nämä mallit ovat tärkeitä. Mainitsit, että ehkä jos sinulta kysyttäisiin niin siellä olisi suomea ainakin 50 %. Ehkä jos ranskalaisilta kysyttäisiin, voi olla, että siellä olisi ranskaa enemmän. Joku voi haluta, että se olisi tosi hyvä koodaamaan. Nyt kielimalleissa se ehkä poikkeaa jonkin verran perinteisistä softakehitysprojekteista, että jokainen mallin paino saattaa vaikuttaa ihan mihin tahansa osa-alueeseen. No okei, on joitain Mixture of Experts -malleja. Mutta joka tapauksessa. Miten nämä päätökset tehdään, mitä siellä mallissa painotetaan? Aiheuttaako tämä jotain sellaista neuvottelua siellä?

(Sampo) Sanotaan, että joo. Aiheuttaa paljon keskustelua. Yritämme löytää, sanotaanko, karkean konsensuksen siitä, miten nämä painotukset menevät. Tämä kysymys on erityisen haastava ehkä juuri niin kuin mainitsit, esim. tällaisilla alueilla kuin koodin kanssa. Koodidataa on saatavilla verrattain paljon. Joten kun meidän laskentaresurssimme ovat kuitenkin tässä vaiheessa vielä jossain määrin tiukempi rajoite, kuin saatavan datan määrä, niin kun katsotaan koodia tai englannin kieltä, niin joudutaan löytämään jonkinlainen kompromissi sen välillä, kuinka paljon painotetaan englannin kieltä ja kuinka paljon painotetaan koodia. Kysymys on vähän yksinkertaisempi pienemmille kielille, kuten Suomen kielille, koska se data on siellä rajoite. Vastaus on kohtuullisen yksinkertainen. Me otamme kaiken mitä meillä on saatavilla, ja se edustaa silti joitain promilleja sitä kokonaisdatamäärää.

(Eric) Joo. Jos mietitään vielä avoimuutta, niin siitä on selkeitä hyötyjä. Se lisää läpinäkyvyyttä, se mahdollistaa sen, että voidaan tutkia syvemmin mitä siellä tapahtuu. Toisaalta ehkä jotkut kriitikot sanovat, että tämä mahdollistaa sen, että kuka tahansa voi ottaa sen mallin painot ja käyttää sitä mihin tahansa vaarallisiin sovelluskohteisiin ilman, että kukaan pystyy kontrolloimaan sitä. Niin miten sinä vertaillet tai mitä ajattelet näistä riskeistä, jotka liittyvät avoimiin malleihin?

(Sampo) Joo, kyllä minä ymmärrän tämän perspektiivin. Jos me olisimme sellaisessa tilanteessa, että me olisimme julkaisemassa avointa mallia, joka on jossain määrin kategorisesti kyvykkäämpi kuin mikään tällä hetkellä muuten saatavilla oleva avoin

malli jollain sellaisella osa-alueella, johon liittyy selkeitä riskejä, esimerkiksi tietoturvallisuus, niin siinä tapauksessa uskon, että meillä olisi hyvinkin erilaisia keskusteluja tällä hetkellä siitä, että kuinka avoimia me voimme olla kaikkein kyvykkäimpien mallien julkaisun yhteydessä. Tässä ehkä tilannetta vähän huvittavasti helpottaa se, että me tulemme Euroopassa akateemisessa toiminnassa ja puhtaasti avoimien mallien kehityksessä selkeästi pikkaisen jäljessä. Joten voidaan aina sanoa, että okei, tätä kapasiteettia ja näitä kykyjä edustavia malleja on ollut jo saatavilla, ja maailma ei ole loppunut.

(Eric) Puhutaan vähän vielä kiinalaisista avoimista malleista. Useat yhdysvaltalaiset pelurit ovat syyttäneet kiinalaisia avointen mallien kehittäjiä näiden isojen yhdysvaltalaismallien ns. tislamisesta, eli siitä, että kiinalaiset mallit opetetaan toistamaan amerikkalaisten mallien vastauksia. Tämä tietysti vastaa siihen datan saatavuusongelmaan. Jos näin tosiaan on, niin onko OpenEuroLLM:n kaltaisten hankkeiden realistista koittaa päihittää näitä kiinalaisia avoimia malleja?

(Sampo) Kyllähän me jossain mielessä olemme kilpailussa. Ja varmasti halutaan, ja uskon, että tulemme olemaan parempia kuin jotkut kiinalaiset mallit. Ehkä en asettaisi tätä ihan niin kuin yhdelle ja samalle linjalle, että tämä on yksinkertainen kilpailu. Meillä on eri painotusalueita, ja pidän hyvinkin todennäköisenä, että meidän mallimme tulevat olemaan vahvempia siinä, miten ne toimivat eurooppalaisilla kielillä, kuin kiinalaiset mallit. Ehkä enemmänkin kuin sitä, että kommentoisin spesifisiä kiinalaisia malleja, niin voisin sanoa, että kielimallien kehityksessä on paljon sellaisia toimijoita, jotka tekevät sellaisia ratkaisuja, oli sitten kyse mallien tislamisesta tai kirjojen polttamisesta tai mistä tahansa muusta, jotka eivät kiinnosta meitä. Eli totta kai se tarkoittaa, että meillä on hieman enemmän haasteita joillain alueilla. Mutta en usko, että nämä ovat kovia rajoitteita. Uskon, että täysin avoimesti tehtävien, täysin avoimiin, julkisesti saataviin materiaaleihin pohjautuvien mallien kehityksessä voidaan tehdä paljon, ja näihin haasteisiin löytyy varmasti ratkaisuja.

(Arno) Eurooppalainen tekoälysuvereniteetti on sellainen slogan, jota kuulee varsinkin poliittisessa keskustelussa aika paljon toistettavan. Mutta onko se reilua edes puhua tällaisesta jonkinlaista teknologisesta suvereniteetista? Vaikka meillä olisikin nämä mallit, niin edelleen se kouluttaminen ja usein käyttäminenkin saattaa riippua hyvin harvoista siruvalmistajista. Ja toisaalta datalähteet ovat usein niin globaaleja, että eivät ne ole puhtaasti eurooppalaisissa käsissä. Niin miten näet tämän suvereniteettikeskustelun? Kuinka kaukana me olemme semmoisesta?

(Sampo) Joo, nämä ovat hyviä pointteja. Tähän liittyy paljon haasteita. Tietenkin raudan puolella me olemme erittäin kaukana. Voidaan sanoa, että emme edes yritä olla kisassa. Tässä mallien kehittämisessä, vaikka totta kai yksittäiset artefaktit, kuten ehkä toivottavasti ne, joita OpenEuroLLM-projekti tulee tuottamaan, niin nämä yksittäiset mallit voivat olla arvokkaita. Mutta näkisin melkein tärkeämpänä sen, että meillä on tämä kapasiteetti tehdä omia malleja. Siihen tietysti liittyy osaaminen, jota me yritämme osaltamme rakentaa. Ja sitten myös se, että meillä on Euroopassa omaa laskentatehoa. Jos katsotaan puhtaasti mallin koulutukseen vaadittavaa laskentatehoa, niin sitä on jo aika paljon. Eli Lumin kaltaiset järjestelmät, nyt Saksassa tuleva

ensimmäinen eurooppalainen exatason laite Jupiter, alkavat olla aika kunnioitettavia supertietokoneita. Tässä laskennassakaan haasteena ei välttämättä ole niinkään se, onko Euroopassa ne GPU-kortit jossain, vaan että halutaanko kollektiivisesti omistaa tästä laskentakapasiteetista tarvittava osa kielimallien tai tekoälymallien kehittämiseen. Jos se tahto on, niin sitten uskon, että raudan puolesta meidän tilanteemme ei ole ollenkaan niin epätoivoinen kuin joskus maalaillaan. Ja viimeiseksi tietysti data. No, tässä haasteet ovat ehkä samat kuin muilla kansainvälisillä toimijoilla. Ei kenelläkään ole omasta puolesta sellaista datamäärää, että pystyttäisiin johtavia malleja kehittämään. Eli toimitaan samoissa rajoitteissa kuin oikeastaan kaikki muutkin.

(Eric) No, puhutaan Sampo lopuksi vielä vähän sinusta. Sinulla on todella vaikuttava julkaisuluettelo. Yli 40 000 viittausta ja h-indeksi on 54, viimeksi kun tarkistin. Olet myös lisäksi työskennellyt monien tosi menestyneiden firmojen kanssa yhteistyössä, kuten esim. Silo AI. Sinulla on papereita Hugging Facen kanssa. Molemmat on ostettu huomattavilla rahasummilla aika hiljattainkin. Sinä työskentelet yliopistotutkijana. Minkä takia tittelisi ei ole esimerkiksi professori Turun yliopistossa tai tutkija jossain yrityspuolella?

(Sampo) Olen kautta urani yrittänyt maksimoida aikaa omistaa tutkimukselle. Tämä on sitten tarkoittanut joitain kompromisseja. Se tarkoitti muun muassa sitä, että väikkärin jälkeen post-dokkailin maailmalla kymmenisen vuotta, koska se antoi niin paljon vapauksia tehdä vain ja ainoastaan tutkimusta. Ei hallintoa, ei opettamista, ei tutkimusrahoituksen tai GPU-laskentarahoituksen hakemista tai muutakaan. No, jossain vaiheessa täytyi hyväksyä jonkinlainen tällaista tiettyä senioriteettia edustavat titteli, ja tämä nyt oli sellainen kompromissi, johon päädyttiin. Vedän pientä tutkimusryhmää ja teen jonkun verran opetusta, mutta ehkä sitten vähän vähemmän näitä muita juttuja. Eli edelleenkin koetan löytää mahdollisimman paljon aikaa tutkimukselle. Ja no, mitä nyt välillä kiroilen supertietokoneille tätä nykyä.

(Eric) Eikös sitä tutkimusta voisi tehdä yrityspuolella myös?

(Sampo) Aivan varmasti, mutta vähän erilaisin rajoittein. Eli omassa työssäni ja meillä laajemmin TurkuNLP:n työssä on aina tämä avoimuus ollut tosi keskeisessä osassa. Ja vaikka esimerkiksi Silo AI:ssa on erittäin kiitettävästi tähän mennessä tehty aidosti avointa, niin sellaisen avoimuuden jatkuvuus ei yrityspuolella ole ikinä täysin taattua. Eli oikeastaan akateeminen maailma on ainoa ympäristö, jossa voin olla aivan varma siitä, että työni on takuulla jatkossakin avoimesti saatavilla.

(Eric) Joo, tosi kunnioitettavaa, ja luulen, että on tärkeää Suomen ja Euroopan kannalta, että meillä on sinunlaisiasi vannoutuneita tutkijoita.

(Arno) Intohimoisia.

(Eric) Intohimoisia.

(Arno) Intohimoisia tutkijoita. Me olemme meidän vierailtamme yleensä kysyneet tämmöistä P(doom)-lukua. Eli todennäköisyys sille, että tämä koko homma menee ihan vihkoon ihmiskunnan kannalta. Niin mikäs olisi sinun P(doom)-lukusi?

(Sampo) Ei, kieltäydyn vastaamasta. Olen todennut, että silloin, kun olen yrittänyt ennustaa tämän alan kehitystä joskus aikaisemmin vähänkään pidemmällä jännitteillä, niin ennustukset ovat menneet pieleen. Eli on luotettavaa historiallista dataa, joka edustaa sitä, että en osaa ennustaa asioita. Tämä nyt on yksi sellainen. No, riippumatta siitä, että kuinka pitkälle tulevaisuuteen pitäisi ennustaa, niin minua paljon fiksummat ihmiset poikkeavat ennusteissaan niin paljon, että joudun toteamaan, että tätä ei todennäköisesti minun lahjoillani voi mielekkäästi yrittää ennustaa. Voin sanoa, että en usko, että puhtaat kielimallit tulevat johtamaan välittömässä lähitulevaisuudessa meitä tuhoon. Mutta pidemmällä jännitteellä en kommentoi.

(Arno) Ehkä hidas dekadenssin ja taantumuksen tie on edessä sitten.

(Eric) Se ei ole sellainen asia, joka valvottaa sinua öisin henkilökohtaisesti?

(Sampo) Ei, eivät nämä pitkäjännitteiset kysymykset. En minä kielimalleja pelkää. Mutta voi olla, että tekoälyn saralla laajemmin tulee joskus jotain, josta kannattaa olla huolissaan. Mutta näen kyllä kielimallit lopulta aika tavallisena koneoppimisteknologiana. Ja niihin liittyvät riskit ovat jotain muuta, kuin että robotit ottavat vallan.

(Arno) Minulla on vielä tällainen kysymys teille molemmille. Koska mielikuvissani edustat nimenomaan esikoulutusta, ja Eric taas jälkikoulutusta, niin mikä tavallaan näiden kahden suhde on? Onko se niin, että jälkikoulutuksessa voidaan fiksata huonosti esikoulutettu malli? Vai miten? Miten tärkeää jälkikoulutuksen kannalta on se esikoulutusvaihe? Ja mitä sitten taas esikoulutuksessa pitää huomioida sen jälkikoulutuksen suhteen?

(Sampo) Vaikea ja laaja kysymys. Jos yrittäisin tätä jotenkin yksinkertaistaa, niin sanoisin ehkä, että esikoulutuksessa yritetään luoda pohjamallille sellaisia jokseenkin ehkä latentteja kykyjä, joita sitten jälkikoulutus voi oikealla tavalla tuoda esiin.

(Eric) Joo, ihan vastaavasti sanoisin tavallaan. Jos ajatellaan esim. näitä päättelymalleja, jotka ratkovat joitain loogisia ongelmia, niin siitä on esimerkiksi sanottu monesti, että jos siellä esikoulutetussa mallissa ei ole mitään kyvykkyyksiä ratkaista sitä mallia, niin ei sitä saada korjattua jälkikoulutuksessa. Mutta jos ajatellaan sitä vastausten jakaumaa, ja se oikea vastaus osuu jonnekin sinne mahdollisiin vastauksiin, niin sitten me saamme ohjattua sen mallin tuottamaan niitä oikeita vastauksia siinä jälkikoulutuksen aikana. Ja vastaavasti ei vaan tällaisia loogisia, mutta myös jos ajatellaan mallien turvallisuutta, niin esikoulutuksessa nyt on varmaan kaikenlaista materiaalia, ja se opettaa paljon kaikkea haitallista. Mutta sitten me saamme jälkikoulutuksessa ohjattua sen sinne turvalliseen ja ihmiselle hyödylliseen käyttäytymiseen. Ehkä sillä lailla tiivistetysti.

(Arno) Eli molempia tarvitaan.

(Sampo) Tottakai.

(Arno) Me haluamme sitten kiittää sinua tästä hyvästä keskustelusta.

(Sampo) Kiitokset.