Kimi Haukka

Kuinka suuret kielimallit oppivat ymmärtämään ja tuottamaan kieltä?

Tietotekniikan Kandidaatintutkielma

5. helmikuuta 2024

Jyväskylän yliopisto

Informaatioteknologian tiedekunta

Tekijä: Kimi Haukka

Yhteystiedot: kimi.k.haukka@student.jyu.fi

Ohjaaja: Tiihonen Timo

Työn nimi: Kuinka suuret kielimallit oppivat ymmärtämään ja tuottamaan kieltä?

Title in English: How do large language models learn and generate?

Työ: Kandidaatintutkielma

Opintosuunta: Tietotekniikan opintosuunta

Sivumäärä: 29+0

Tiivistelmä: Tekoälyn ja koneoppimisen, erityisesti transformer-pohjaisten kielimallien, kehitys on mullistanut kieltenkäsittelyn. Tässä tutkielmassa tarkastelemme näiden mallien kykyä tuottaa ja ymmärtää kieltä, keskittyen niiden oppimisprosesseihin ja kielen rakenteiden sisäistämiseen. Tutkimme kuinka transformer-mallien 'self-attention-mekanismi' edistää tekstin syvällistä ymmärrystä ja kuinka nämä mallit kehittävät kykyä ennustaa tulevia sanoja ja lauseita, mikä auttaa hahmottamaan kieliopillisia ja semanttisia rakenteita paremmin.

Tutkielmassa käsitellään myös, missä määrin näiden mallien kielen ymmärtäminen on aitoa ja missä määrin se perustuu vaikutelman luomiseen. Vaikka mallit ovat kehittyneitä, niiden kyky ymmärtää kieltä ei ole yhtä syvällinen kuin ihmisen kyky ymmärtää kieltä. Tutkielmassa korostetaan, että näiden mallien todellinen ymmärryskyky jää rajoitetuksi, vaikka ne ovatkin kehittyneet tuottamaan tekstiä, joka vaikuttaa ymmärtävän kieltä.

Avainsanat: Tekoäly, LLM, NLP, Transformer-arkkitehtuuri

Abstract: The development of artificial intelligence and machine learning, especially transformer-based language models, has revolutionized language processing. In this study, we examine the ability of these models to generate and understand language, focusing on their learning processes and the internalization of language structures. We explore how the 'self-attention mechanism' of transformer models contributes to a deep understanding of text and how the-

se models develop the ability to predict future words and sentences, which helps in grasping grammatical and semantic structures better.

The study also addresses the extent to which the language understanding of these models is genuine and the extent to which it is based on creating an impression. Although the models are advanced, their ability to understand language is not as profound as the human ability to understand language. The study emphasizes that the real comprehension capability of these models remains limited, even though they have developed to produce text that appears to understand language.

Keywords: Artificial intelligence, LLM, NLP, Transformer architecture

Kuviot

Taulukot

Sisällys

1 Johdanto

Tietojenkäsittelyn ala on kokenut merkittäviä muutoksia viime vuosina. Aikaisemmin ala keskittyi pääasiassa yksinkertaisiin laskentaoperaatioihin, mutta nyt se on siirtynyt näistä eksplisiittisesti ohjelmoiduista toiminnoista kohti aineistosta oppivia malleja ja sääntöjä. Tämä siirtyminen on mahdollistanut tekoälyn ja koneoppimisen sovellusten edistyneiden ominaisuuksien kehittymisen. Tämä muutos on vaikuttanut syvästi moniin teollisuudenaloihin, muokaten sekä työtapojamme että käsityksiämme datan merkityksestä ja potentiaalista.

Koneoppiminen ja erityisesti luonnollisen kielen käsittely (NLP) ovat saavuttaneet huomattavia edistysaskelia viime vuosina. Kehittyneet suuret kielimallit edustavat luonnollisen kielenkäsittelyn nykyistä huippua, osoittaen merkittävää kehitystä yksinkertaisista järjestelmistä monimutkaisiin malleihin. Nämä mallit eivät ainoastaan tuota yhtenäistä tekstiä vaan kykenevät myös vastaamaan kysymyksiin ja luomaan luovaa sisältöä. Herää kuitenkin kysymys: ymmärtävätkö nämä mallit todella kieltä ja kontekstia, vai luovatko ne vain vaikutelman ymmärtämisestä?

Tämä viimeinen kysymys on keskiössä tässä tutkielmassa. Mitkä ovat keskeiset mekanismit ja rakenteet, jotka mahdollistavat kielen tuottamisen ja ymmärtämisen näissä malleissa? Vaikutelma kielen ja kontekstin ymmärtämisestä on vahva, mutta mihin se perustuu? Miltä osin se on pelkkää vaikutelmaa ja miltä osin sitä voidaan kutsua ”aidoksi” ymmärtämiseksi? Tämän tutkielman tavoitteena on syventyä näihin kysymyksiin ja ymmärtää paremmin suurten kielimallien oppimisprosesseja sekä niitä mekanismeja ja rakenteita, jotka tekevät tämän mahdolliseksi.

Tässä tutkielmassa tarkastellaan tekoälyn ja kielimallien alueiden neljää keskeistä aihetta. Toinen luku tutkii tekoälyn perusteita, syventyen erityisesti syväoppimisen merkitykseen ja sen vaikutuksiin tekoälyn kehityksessä. Kolmannessa luvussa tutkitaan kielimallien kehitystä, alkaen historiallisista malleista ja edeten kohti nykypäivän monimutkaisia neuroverkkoja ja transformer-arkkitehtuureja. Neljäs luku pureutuu kielimallien oppimisprosesseihin ja kielenkäsittelyn haasteisiin, tarkastellen näiden mallien kyvykkyyksiä ja rajoitteita kielen simuloinnissa. Viides luku keskittyy ”aidon ymmärryksen” konseptiin tekoälyssä, vertaillen

sitä inhimilliseen kognitioon ja pohtien, miten tilastollinen analyysi vaikuttaa näiden mallien toimintaan.

2 Tekoäly

Tämä luku syventyy tekoälyn keskeisiin käsitteisiin ja sovelluksiin, tarkastellen sen teoreettisia perusteita sekä käytännön toteutuksia. Aloitamme määrittelemällä, mitä tekoäly itse asiassa on, ja miten se vaikuttaa nykypäivän teknologiaan. Keskitymme erityisesti koneoppimiseen, joka on tekoälyn ydinalue, ja käymme läpi sen päämuodot: valvotun ja valvomattoman oppimisen. Käsittelyssä korostuu myös syväoppimisen merkitys, jossa uusimmat innovaatiot, kuten GPT (Generative Pre-training Transformer), näyttelevät keskeistä roolia.

2.1 Mikä on tekoäly?

Russellin ja Norvigin (2020) mukaan tekoäly on tietokone tai tietokoneohjelma, joka on suunniteltu jäljittelemään ja suorittamaan tehtäviä, jotka yleensä vaativat ihmisen älykkyyttä. Tämän määritelmän keskeisenä piirteenä on tekoälyn kyky suorittaa tehtäviä, jotka vaativat päätöksentekoa ja oppimista ilman yksityiskohtaista ihmisen ohjausta, sisältäen tehtäviä kuten puheen tunnistamisen, pelien pelaamisen, kuvioiden tunnistamisen ja autolla ajamisen. Tekoäly oppii näitä tehtäviä prosessoimalla suuria määriä dataa, etsimällä malleja ja soveltamalla niitä omassa päätöksenteossaan. Vaikka tekoälyjärjestelmiä usein ohjataan ja valvotaan ihmisten toimesta, jotkin tekoälyjärjestelmät kykenevät oppimaan itsenäisesti ilman ihmisen aktiivista ohjausta, esimerkiksi pelaamalla videopeliä toistuvasti ja oppimalla pelin säännöt ja strategiat. Tämä itsenäinen oppimiskyky on keskeinen osa tekoälyn kehitystä ja sen soveltamista eri alueilla (Russell ja Norvig 2020). Lisäksi, tekoälyjärjestelmien kyky prosessoida suuria datamääriä ja tunnistaa malleja on merkittävä edistysaskel niiden suorituskyvyssä, erityisesti monimutkaisten ongelmien ratkaisemisessa ja päätöksenteossa (Mitchell 1997).

Inhimillinen ajattelu	**Rationaalinen ajattelu**
”Jännittävä uusi yritys saada tietokoneet ajattelemaan... *machines with minds*, täydessä ja kirjaimellisessa merkityksessä.” (Haugeland 1985)	”Mentaalisten kykyjen tutkiminen laskennallisten mallien avulla.” (Charniak ja McDermott 1985)
”[Automatisoidut] toiminnot, jotka yhdistämme ihmisen ajatteluun, toiminnot kuten päätöksenteko, ongelmanratkaisu, oppiminen...” (Bellman 1978)	”Tutkimus laskelmista, jotka mahdollistavat havaitsemisen, päättelyn ja toiminnan.” (Winston 1992)
Inhimillinen toiminta	**Rationaalinen toiminta**
”Taidetta luoda koneita, jotka suorittavat toimintoja, jotka vaativat älykkyyttä, kun ihmiset käyttävät niitä.” (Kurzweil 1990)	”Laskennallinen älykkyys on tutkimus älykkäiden agenttien suunnittelusta.” (Poole, Mackworth ja Goebel 1998)
”Tutkimus siitä, miten saada tietokoneet tekemään asioita, joissa ihmiset ovat tällä hetkellä parempia.” (Rich ja Knight 1991)	”AI ... on huolissaan artefaktien älykkäästä käyttäytymisestä.” (Nilsson 1998)

Taulukko 1. Tekoälyn määritelmiä, järjestettynä neljään kategoriaan. Taulukko on suomennettu alkuperäisestä lähteestä: Russell ja Norvig, *’Artificial Intelligence: A Modern Approach’* (2020).

2.2 Koneoppiminen

Koneoppiminen on tekoälyn dynaaminen ja välttämätön alue, joka muuttaa tapaa, jolla tietokonejärjestelmät omaksuvat ja soveltavat tietoa ilman eksplisiittistä ohjelmointia. Tämän alueen keskiössä on tilastollisten menetelmien ja algoritmien soveltaminen, mikä antaa koneille mahdollisuuden oppia aikaisemmista kokemuksista ja datan käsittelystä. Tässä yhteydessä on verrattava koneoppimista perinteisiin asiantuntijajärjestelmiin, jotka taas päinvastoin hyödyntävät ekspliittisesti syötettyä tietoa (Aniba ym. 2008). Koneoppiminen mahdollistaa suurten tietomäärien tulkinnan sekä vaativien tehtävien ratkaisemisen, kuten lääketieteellisen datan käsittely (Murphy 2012, s. 29).

Koneoppiminen jakautuu kahteen päähaaraan: valvottuun ja valvomattomaan oppimiseen (Kapitanova ja Son 2012, s. 5). Valvotussa oppimisessa mallit rakennetaan käyttäen etukäteen määriteltyjä, merkittyjä tietoaineistoja, jotka opettavat koneille haluttuja malleja ja ennustuskykyä. Esimerkiksi, roskapostin tunnistamisessa käytetään valvottua oppimista opettaakseen järjestelmiä erottamaan hyödylliset sähköpostit ei-toivotuista (Goldberg 2017, s. 35). Valvomattomassa oppimisessa, toisaalta, koneet työskentelevät merkitsemättömien datajoukkojen kanssa, pyrkien löytämään itsestään piileviä rakenteita ja yhteyksiä. Tällaisia sovelluksia nähdään esimerkiksi markkinatutkimuksessa, jossa yritykset ryhmittelevät asiakkaita heidän käyttäytymisensä perusteella ilman ennalta määriteltyjä luokkia. Nämä koneoppimisen metodit ovat keskeisiä tekoälyn edistymisessä, tarjoten tietokoneille mahdollisuuden itsenäiseen sopeutumiseen ja oppimiseen. Erityisesti syväoppimisen alueella, jossa käytetään monimutkaisia neuroverkkoja, koneoppimisen merkitys korostuu entisestään. Neuroverkkojen avulla koneoppiminen ei ainoastaan käsittele suuria datamääriä, vaan myös oppii tunnistamaan ja toistamaan monimutkaisia kuvioita ja suhteita datassa (Murphy 2012). Neuroverkkoja käsitellään tarkemmin luvussa 2.3.

2.3 Neuroverkkojen perusteet ja toimintaperiaatteet

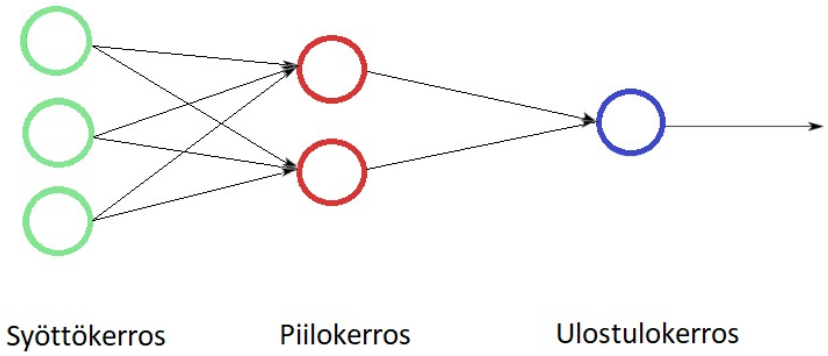
Neuroverkot ovat koneoppimisen ja tekoälyn keskeinen osa-alue, joka jäljittelee ihmisen aivojen toimintaa informaation prosessoinnissa. Perusyksikkönä neuroverkoissa toimii neuroni, joka vastaanottaa syötteitä, prosessoi niitä ja tuottaa tulosteita. Nämä keinotekoiset neuronit on järjestetty kerroksiin, joista jokainen suorittaa tietynlaisen datan muunnoksen (Goodfellow, Bengio ja Courville 2016).

Neuroverkon perusrakenne koostuu kolmesta päätyyppisestä kerroksesta: syöttökerros (input layer), piilokerrokset (hidden layers) ja ulostulokerros (output layer). Syöttökerros vastaanottaa raakadatan, piilokerrokset suorittavat monimutkaisia laskentoja datan attribuuttien oppimiseksi ja ulostulokerros tuottaa lopullisen tuloksen, esimerkiksi luokittelun tai ennusteen (LeCun, Bengio ja Hinton 2015).

Neuroverkkojen oppimiskyky perustuu siihen, että ne voivat säätää neuronien välisiä yhteyksiä, jotka tunnetaan painoina (weights). Oppimisprosessi neuroverkoissa tapahtuu virheen

takaisinvirtauksen (backpropagation) ja gradientin laskun (gradient descent) avulla. Tämä prosessi sisältää virheen laskemisen ennusteen ja todellisen tuloksen välillä sekä painojen päivittämisen virheen minimoimiseksi (Rumelhart, Hinton ja Williams 1986).

Neuroverkot ovat osoittautuneet tehokkaiksi monilla alueilla, kuten kuvan- ja puheentunnistuksessa, luonnollisen kielenkäsittelyssä sekä monimutkaisten ennustemallien luomisessa. Niiden monipuolisuus ja kyky oppia monenlaisista dataformaateista tekevät niistä yhden tärkeimmistä työkaluista nykyaikaisessa tekoälyssä.

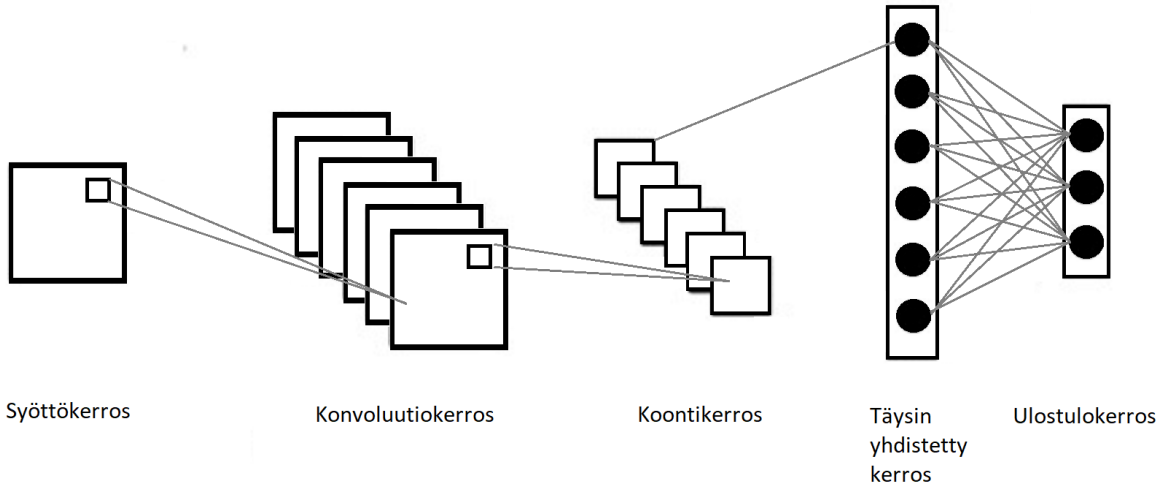


Kuvio 1. Yksinkertainen esitys kolmikerroksisesta neuroverkosta. Neuroverkossa on syöttökerros, jossa on kolme solmua, yksi piilokerros, jossa on kaksi solmua, ja ulostulokerros, jossa on yksi solmu.

2.4 Syväoppiminen

Syväoppiminen, yhtenä koneoppimisen innovatiivisimmista ja kiehtovimmista ala-alueista, hyödyntää neuroverkkoarkkitehtuureja, jotka ovat saaneet innoituksensa ihmisaivojen toiminnasta. Tässä kontekstissa neuroverkot, monikerroksisina rakenteina, tarjoavat syvällisen lähestymistavan datan käsittelyyn (Goodfellow, Bengio ja Courville 2016). Tämän prosessin ydin on neuroverkkojen kyvyssä oppia itse, havaita malleja ja tehdä päätelmiä suurista ja monimuotoisista datamääristä. Tämä ominaisuus on ratkaisevan tärkeä, kun käsitellään monimutkaisia kielillisiä ja visuaalisia tehtäviä, kuten kielenkäsittelyä tai kuvien analysointia. Syväoppimisen todelliset sovellukset ovat monipuolisia ja vaikuttavia. Esimerkiksi konvoluutioneuroverkkojen käyttö kuvantunnistuksessa on mullistanut, kuinka koneet näkevät ja tulkitsevat visuaalista tietoa, tuoden uusia ulottuvuuksia lääketieteelliseen diagnostiikkaan

ja digitaaliseen kuvankäsittelyyn. Puheentunnistuksessa ja chatbot-teknologioissa syväoppiminen mahdollistaa koneiden ja ihmisten välisen sujuvan ja luonnollisen kommunikaation, mukaillen ihmisen puhetapaa ja ymmärrystä (Goodfellow, Bengio ja Courville 2016).



Kuvio 2. Kuvassa on konvoluutioneuroverkon perusrakenne, joka koostuu syöttökerroksesta, konvoluutio- ja koontikerroksista piirteiden erottelemiseksi, täysin yhdistetystä kerroksesta päättelyyn ja ulostulokerroksesta lopputuloksen näyttämiseen.

Luonnollisen kielenkäsittelyn alueella syväoppiminen on avannut ovet ennennäkemättömälle tekstigeneraation kyvylle, jossa koneet pystyvät tuottamaan tekstiä, joka on niin sulavaa ja yhtenäistä, että se muistuttaa ihmisen kirjoittamaa (Brown ym. 2020). Tämä monitasoinen oppimisprosessi, joka jäljittelee ihmisaivojen toimintatapaa, on syväoppimisen keskiössä. Neuroverkkojen kerrostunut rakenne mahdollistaa datan käsittelyn ja oppimisen useilla eri tasoilla, mikä johtaa tarkempaan analyysiin ja suorituskykyyn monimutkaisten tehtävien parissa. Tämän alueen tärkeys ja tehokkuus koneoppimisessa on selkeästi dokumentoitu teoksessa ”Deep Learning” (Goodfellow, Bengio ja Courville 2016).

3 Kielimallinnus ja sen kehitys

Tässä luvussa perehdytään luonnollisten kielimallien evoluutioon ja niiden merkitykseen tekoälyssä. Alkaen perusteista, tämä luku selventää, kuinka nämä mallit ovat siirtyneet alkuaikojen yksinkertaisista algoritmeista nykypäivän edistyneisiin järjestelmiin, jotka hallitsevat ihmiskielen monimuotoisuutta. Tarkastelemme ensin, miten n-grammimallit loivat perustan kielelliselle analyysille, mutta törmäsivät rajoituksiinsa kielen syvällisemmässä käsittelyssä. Luvussa perehdytään myös luonnollisten kielimallien laajaan soveltuvuuteen ja niiden rooliin teknologian eri alueilla, aina chatboteista ääniassistentteihin. Erikoistarkastelussa ovat myös neuroverkot ja niiden keskeinen asema kielellisissä tehtävissä, erityisesti transformer-arkkitehtuurin innovaatiot ja niiden vaikutukset kielimallinnukseen. Lopuksi luku pohtii suurten kielimallien, kuten GPT:n (Generative pre-training transformer), vaikutusta ja esittelee niiden sovelluksia sekä tärkeitä eettisiä pohdintoja niiden käytössä.

3.1 Luonnollisten kielimallien perusteet

Luonnolliset kielimallit ovat algoritmeja tai ohjelmia, jotka on kehitetty tunnistamaan, ymmärtämään, käsittelemään ja tuottamaan ihmiskieltä. Nämä kyseiset kielimallit ovat keskeisiä tekoälyssä ja koneoppimisessa, sillä ne mahdollistavat ihmisen ja koneen välisen mahdollisimman luontaisen kommunikaation (Jurafsky ja Martin 2019). Näiden kielimallien kehittäminen on ollut yksi koneoppimisen ja tekoälyn tutkimuksen keskeisistä tavoitteista. Nykyaikaiset kielimallit perustuvat neuroverkkoihin, jotka mahdollistavat oppimisen ja mukautumisen erilaisiin kielenkäsittelyn tilanteisiin. Tällaiset mallit käyttävät monimutkaisia matemaattisia ja tilastollisia menetelmiä, kuten self-attention-mekanismeja ja sijaintikoodauksia, kielen syvällisempään ymmärtämiseen ja tuottamiseen (Naveed ym. 2023).

Luonnolliset kielimallit voivat olla yksinkertaisia, perustuen selkeisiin sääntöihin ja malleihin, tai monimutkaisempia, kuten nykyaikaiset neuroverkkoihin pohjautuvat mallit, jotka kykenevät oppimaan ja mukautumaan uusiin kielenkäyttötilanteisiin. Varhaiset mallit keskittyivät yksinkertaisiin tekstianalyysin muotoihin, kun taas nykyaikaiset mallit, kuten transformer-arkkitehtuuria hyödyntävät järjestelmät, ovat kyenneet luomaan monimutkaisia

ja yhtenäisiä tekstejä sekä ymmärtämään kieltä syvällisemmällä tasolla (Jurafsky ja Martin 2019). Transformer-arkkitehtuurin yksityiskohtaisempi tarkastelu ja sen vallankumouksselliset vaikutukset kielimallinnuksen alalla esitetään tarkemmin luvussa 3.4.

3.2 Ensimmäiset luonnolliset kielimallit ja niiden rajoitukset

Kun katsomme taaksepäin luonnollisten kielimallien historiassa, n-grammimallit nousevat esiin ensimmäisinä merkittävinä askelina kielen prosessoinnin ja ymmärtämisen automatisoinnissa (Jurafsky ja Martin 2019, s. 56). Näiden mallien juuret ovat tilastollisessa analyysissä, ja ne kehitettiin alun perin tehtäviin, kuten automaattiseen tekstinkäsittelyyn ja kielen rakenteen mallintamiseen. N-grammimalli käyttää yksinkertaista lähestymistapaa, jossa tarkastellaan sanojen sekvenssejä (esimerkiksi kolmen sanan trigrammeja) ennustaakseen seuraavan sanan todennäköisyyksiä. Otetaan esimerkiksi lause "Kissat nukkuvat päivisin kylmällä lattialla". Tässä trigrammimalli voi jakaa lauseen osiin kuten "Kissat nukkuvat päivisin", "nukkuvat päivisin kylmällä"ja "päivisin kylmällä lattialla", joista malli oppii näiden yhdistelmien yleisyyttä ja käyttöä kielidatassa (Jurafsky ja Martin 2019, s. 32).

Kissat nukkuvat päivisin kylmällä lattialla

1-Gram	2-Gram	3-Gram
Kissat	Kissat nukkuvat	Kissat nukkuvat päivisin
nukkuvat	nukkuvat päivisin	nukkuvat päivisin kylmällä
päivisin	päivisin kylmällä	päivisin kylmällä lattialla
kylmällä	kylmällä lattialla	
lattialla		

Kuvio 3. Kuvassa on yksinkertainen esimerkki n-grammimallien toiminnasta, jossa 1-grammimalli, 2-grammimalli ja 3-grammimalli ennustavat seuraavia sanoja lauseessa ’Kissat nukkuvat päivisin kylmällä lattialla’.

Kuten (Jurafsky ja Martin 2019) totesivat, vaikka nämä mallit olivat hyödyllisiä sanojen esiintymisen ja peruskielirakenteiden ennustamisessa, niillä oli olennaisia puutteita. Erityisesti ne eivät kyenneet käsittelemään kielen pitkän aikavälin riippuvuuksia ja monimutkai-

sempia rakenteita. Tämä johtui siitä, että n-grammimallit rajoittuvat tarkastelemaan vain pieniä, rajattuja sanojen ketjuja, jolloin laajempi konteksti jäi huomiotta. Tämä rajoitus heijastui mallien kyvyssä ymmärtää ja tuottaa kieltä, joka muistuttaa aitoa ihmisen käyttämää kieltä. Vaikka n-grammimallit eivät ole yhtä monimutkaisia kuin viimeisimmät neuroverkkojen avulla toteutetut kielimallit, kuten rekurrentit neuroverkot (RNN:t) ja transformer-mallit – joita käsitellään tarkemmin luvussa 3.4 – ne ovat silti keskeisiä peruskäsitteiden ymmärtämisen kannalta kielen mallintamisessa (Jurafsky ja Martin 2019).

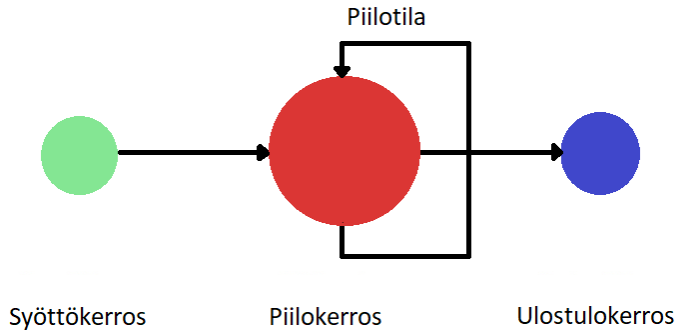
3.3 Luonnollisten kielimallien merkitys teknologiassa

Luonnolliset kielimallit ovat kehittyneet olemaan yksi koneoppimisen ja tekoälyn keskeisimmistä työkaluista, joita hyödynnetään laajasti monissa sovellusalueissa. Ne ovat olennaisia monimutkaisten kielellisten tehtävien, kuten käännöspalveluiden, chatbottien, ääniassistenttien, ja automaattisen tekstianalyysin toteuttamisessa. Luonnolliset kielimallit edistävät kielitieteen tutkimusta tarjoamalla syvällistä ymmärrystä kielen rakenteesta ja käytöstä (Jurafsky ja Martin 2019). Teollisuudessa ne auttavat parantamaan asiakaskokemusta ja tehostamaan liiketoiminnan prosesseja, esimerkiksi automatisoimalla asiakaspalvelua ja analysoimalla asiakaspalautetta (Hirschberg ja Manning 2015). Tämän lisäksi, luonnolliset kielimallit tukevat myös kognitiotieteitä ja kieliteknologiaa, tarjoten uusia näkökulmia ihmisen kielenkäytön ymmärtämiseen ja simulointiin (Goodfellow, Bengio ja Courville 2016).

3.4 Transformer-pohjaiset mallit ja niiden vallankumouksellisuus

Transformer-arkkitehtuuri, esitelty ensimmäisen kerran vuonna 2017 merkittävässä tutkimusartikkelissa *”Attention Is All You Need”* (Vaswani ym. 2017), on edustanut merkittävää edistysaskelta kielimallinnuksen saralla. Keskeinen osa tätä arkkitehtuuria on sen ”self-attention-mekanismi”, joka erottuu kyvyllään vertailla ja yhdistellä input-sarjan kaikkia osia keskenään. Tämä on huomattava ero verrattuna perinteisiin sekvenssimallinnusmenetelmiin, kuten rekurrentteihin neuroverkkoihin (RNN) ja pitkän aikavälin muistiin (LSTM), jotka prosessoivat tietoa sekvensseissä peräkkäisesti (Bengio, Simard ja Frasconi 1994). Toisin kuin transformerit, rekurrentit neuroverkot käsittelevät tietoa lineaarisesti, mikä voi aiheut-

taa ongelmia pitkien riippuvuuksien ymmärtämisessä. Pitkän aikavälin muistit parantavat tätä prosessointia ”porttien” avulla, jotka säilyttävät ja unohtavat tietoa tehokkaammin, mutta nekin kohtaavat haasteita erittäin pitkissä sekvensseissä (LeCun, Bengio ja Hinton 2015).

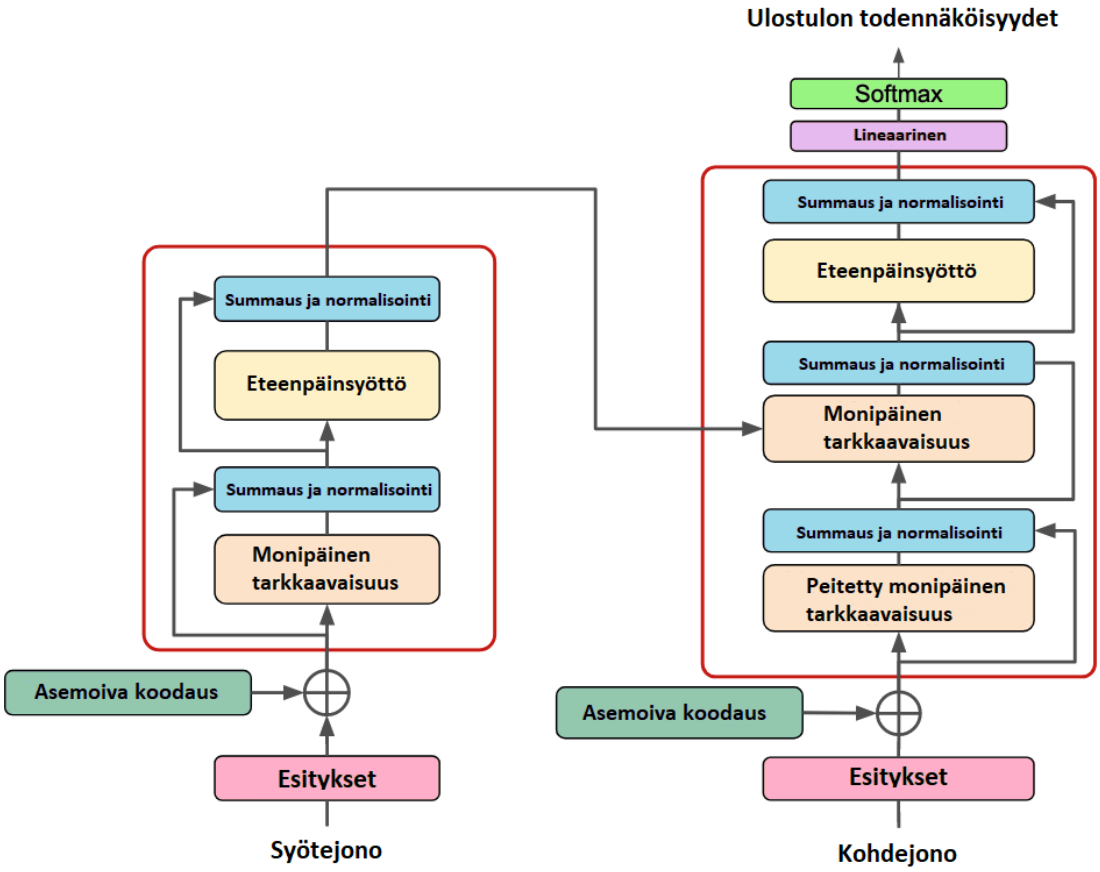


Kuvio 4. Rekurrentin neuroverkon (RNN) malli, jossa on syöttökerros, piilokerros ja ulostulokerros. Lisäksi piilotila, joka säilyttää ja hyödyntää aikaisempien syötteiden tietoja myöhempää prosessointia varten.

Self-attention-mekanismi mahdollistaa mallille keskittymisen olennaisiin tietoihin ja suhteisiin, parantaen kielen ymmärtämistä ja kontekstin hallintaa. Self-attention-mekanismi on keskeinen osa transformer-arkkitehtuuria, joka erottuu perinteisistä sekvenssimallinnusmenetelmistä, kuten n-grammeista. Toisin kuin n-grammit, jotka rajoittuvat käsittelemään kielen rakenneosia rajatussa aikavälissä ja ovat siten haavoittuvaisia pitkien riippuvuuksien ymmärtämisessä, self-attention mahdollistaa transformer-mallien, kuten BERT (Bidirectional Encoder Representations from Transformers) ja GPT (Generative Pre-trained Transformer), keskittyä kielellisten ja monimutkaisten yhteyksien ymmärtämiseen. BERTin läpimurto, kuten Devlin et al. (2018) ovat kuvailleet, on sen kyky analysoida tekstiä kaksisuuntaisesti, tarjoten parannettua ymmärrystä kielellisistä malleista. Tämä kaksisuuntainen analyysi tarkoittaa, että BERT tutkii sekä sanoja, jotka tulevat ennen käsiteltävää sanaa että sanoja, jotka tulevat sen jälkeen. Näin se saa kokonaiskuvan siitä, miten sana liittyy ympäröivään tekstiin, auttaen ymmärtämään sen merkitystä paremmin. Radford et al. (2019) tutkimuksessa esitellään GPT-mallit, jotka hyödyntävät puolestaan syväoppimista generoidakseen yhtenäisiä ja luovia tekstejä, mikä on avannut uusia sovellusmahdollisuuksia luovassa kirjoittamisessa ja dialogijärjestelmissä.

Näiden innovaatioiden myötä tekoäly ja koneoppiminen ovat astuneet uudelle tasolle luon-

nollisen kielenkäsittelyssä. Tämän teknologian avulla tutkijat ja kehittäjät voivat tutkia kieltä ja sen ilmiöitä aikaisempaa syvällisemmin, avaten uusia ovia kielen tutkimukselle ja sovelluksille. Transformer-arkkitehtuurin esittely on ollut merkittävä virstanpylväs, joka on muokannut käsitystämme siitä, mihin koneoppiminen ja tekoäly voivat venyttää rajojaan kielellisten tehtävien parissa.



Kuvio 5. Tämä kuva esittää transformer-arkkitehtuurin rakennetta, joka sisältää syöte- ja kohdejonot, asemoivan koodauksen sekä monipäisen tarkkaavaisuuden komponentit. Prosessin lopussa Softmax- ja lineaariset kerrokset tuottavat lopputuloksen todennäköisyydet.

4 Kielimallit ja kielenkäsittelyn haasteet

Tässä luvussa syvennymme kielimallien ja transformer-arkkitehtuurien, oppimisprosesseihin sekä kielenkäsittelyn haasteisiin. Tarkastelemme huolellisesti näiden mallien oppimismenetelmiä ja niihin liittyviä teknologisia innovaatioita. Käsittelemme myös kielimallien koulutusprosessia, korostaen datan monimuotoisuuden ja koulutusaineiston valinnan merkittävää roolia. Lopuksi analysoimme kielimallien kykyjä ja rajoituksia simuloida ihmisen kielen ymmärrystä, tuoden esille sekä niiden potentiaalin että kohtaamamme haasteet.

4.1 Suuret kielimallit

Suuret kielimallit (engl. Large language model), kuten GPT ja BERT, ovat vallankumouksellinen näyttö siitä, miten pitkälle tekoäly on edennyt kielen ymmärtämisessä ja tuottamisessa. Nämä mallit hyödyntävät syväoppimisen menetelmiä, kuten transformer-arkkitehtuuria, joka mahdollistaa kontekstin tehokkaan hyödyntämisen ja pitkien riippuvuuksien käsittelyn kielenkäsittelyssä (Vaswani ym. 2017). Esimerkiksi GPT-mallien toiminta perustuu generatiiviseen koulutukseen, jossa malli oppii ennustamaan tekstiä suuresta, monipuolisesta datamäärästä. Tämä tekee niistä erityisen vahvoja luomaan tekstiä, joka heijastaa syvällistä kontekstuaalista ymmärrystä ja kielen vivahteita (Brown ym. 2020). Suurten kielimallien kyky käsittää ja soveltaa tietoa monipuolisissa ympäristöissä – kuten vastata kysymyksiin, luoda tarinoita ja jopa generoida koodia – avaa uusia mahdollisuuksia tekoälyn soveltamiseen (Brown ym. 2020).

Suurten kielimallien kehitys on myös mahdollistanut niiden käytön opetuksessa ja oppimisessa. Koneet voivat nyt toimia virtuaalisina opettajina, jotka tarjoavat räätälöityä sisältöä ja auttavat kielen oppimisessa, tarjoten esimerkkejä, selityksiä ja harjoituksia reaaliaikaisesti (Bommasani ym. 2021). Lisäksi, nämä mallit ovat osoittaneet kykynsä toimia kieliteknologian tutkimuksen apuvälineinä, antaen tutkijoille mahdollisuuden testata kieliteorian hypoteeseja tai arvioida kielen eri ilmiöitä mittavien datamassojen avulla.

Samalla, suurten kielimallien eettinen käyttö on kuitenkin herättänyt keskustelua, sillä niiden kyky generoida vakuuttavaa tekstiä voi johtaa väärinkäyttöön, kuten disinformaation levittä-

miseen. Tämä korostaa tarvetta vastuulliseen käytäntöön ja menetelmiin, jotka varmistavat generoidun tiedon luotettavuuden ja tarkkuuden (Weidinger ym. 2021).

Suurten kielimallien kehitys merkitsee merkittävää edistysaskelta koneoppimisen ja tekoälyn kentällä. Niiden tarjoamat mahdollisuudet ja haasteet muokkaavat jatkuvasti tieteen ja teknologian rajoja, avaten uusia näköaloja soveltamisen ja tutkimuksen maailmassa.

4.2 Miten suuret kielimallit oppivat ymmärtämään kieltä?

Suurten kielimallien ymmärrys kielen rakenteista ja merkityksistä juontaa juurensa niiden kyvystä analysoida laajoja tekstiaineistoja. Nämä mallit, hyödyntäen syväoppimista ja erityisesti transformer-arkkitehtuuria, tunnistavat ja mallintavat kielen moninaisuutta tilastollisten menetelmien kautta. Bengio ja hänen kollegansa (Bengio ja Rejan Ducharme 2003) ovat korostaneet, miten koneoppimismenetelmät, varsinkin syväoppiminen, oppivat hahmottamaan ja tulkitsemaan monimutkaisia tietoaineistoja. Tämä luo perustan, jolla kielimallit voivat tehostaa kielellistä prosessointiaan.

Transformer-arkkitehtuurin keskiössä on sen kyky korostaa tärkeimpiä osia tekstistä ”self-attention-mekanismin” avulla. Tämä erottaa transformerit aiemmista malleista, kuten rekurrenteista neuroverkoista, jotka käsittelevät tietoa lineaarisesti. Vaswani ja työryhmä (Vaswani ym. 2017) kuvailevat, kuinka tämä mekanismi mahdollistaa tekstikokonaisuuksien syvällisemmän ymmärtämisen.

Toisaalta, ennakoivat mallit kuten GPT oppivat ennustamaan seuraavia sanoja tai lauseita suurista tekstikokoelmista. Radford ja hänen tiiminsä (Radford ym. 2019) ovat osoittaneet, kuinka tällainen ennustava mallintaminen syventää käsitystämme kieliopillisista ja semanttisista rakenteista.

Suurten kielimallien oppimisprosessi perustuu myös suurten datamäärien hyödyntämiseen. Niin kutsuttu ”unsupervised learning” eli valvomaton oppiminen, jossa mallit prosessoivat valtavia määriä merkitsemätöntä tekstiä, on osoittautunut tehokkaaksi tavaksi kehittää malleja, jotka ymmärtävät ja tuottavat ihmiskieltä autenttisesti (Devlin ym. 2018). Tämä menetelmä mahdollistaa mallien kehittämisen, jotka eivät vain toista sanoja ja fraaseja, vaan myös

ymmärtävät niiden merkityksen laajemmassa kontekstissa.

4.3 Kielimallien tehokas koulutus ja koulutusdatan laadun tärkeys

Kielimallien tehokkuus kielen prosessoinnissa ja tuottamisessa pohjautuu niiden koulutukseen, joka tapahtuu laajojen ja monimuotoisten tekstiaineistojen avulla. Mallien jatkuva altistuminen erilaisille datamassoille mahdollistaa niiden sopeutumisen kielellisiin rakenteisiin ja niiden sisältämiin merkityksiin. Tämä oppimisprosessi on erittäin tärkeä, kun tavoitteena on kehittää malleja, jotka kykenevät hallitsemaan kielen moninaisuutta ja monimutkaisuutta. Viimeaikaisissa tutkimuksissa, kuten Xie et al. (2023) on todettu, että tarkoin valittu koulutusdata on avainasemassa sekä yleiskäyttöisten että erikoistuneiden kielimallien, kuten GPT:n, tehokkuudessa. Heidän tutkimuksensa korostaa painotetun uudelleenotannan (importance resampling) merkitystä datan valinnassa, mikä on osoittautunut hyödylliseksi lähestymistavaksi tekstidatan monimuotoisen tilan hallinnassa (Xie ym. 2023). Painotettu uudelleenotanta tarkoittaa tilastollista menetelmää, jossa datan osia arvostetaan eri tavoin näytteenoton yhteydessä, antaen lisäpainoarvoa tietyille datan osille niiden merkityksen tai harvinaisuuden perusteella, mikä edistää kielimallin kykyä oppia ja sopeutua kielellisiin erityispiirteisiin ja monimuotoisuuteen.

Koulutusdatan laadun ja monimuotoisuuden rooli on erityisen merkittävä. Tutkijat kuten Hovy ja Spruit ovat painottaneet monipuolisen datan merkitystä ennakkoluulojen vähentämisessä ja mallien yleistettävyyden parantamisessa. Heidän mukaansa on tärkeää sisällyttää koulutusaineistoon eri kielimuotoja ja -variantteja, jotta voidaan edustaa kielen moninaisuutta ja vähentää kulttuurista vinoutumista (Hovy ja Spruit 2016).

Lisäksi, kielimallien eettinen ja vastuullinen käyttö on noussut keskeiseksi huolenaiheeksi. Tutkijat Bender ja kollegat ovat ehdottaneet, että kielimallien koulutuksessa tulisi ottaa huomioon eri kulttuurien ja kielten erityispiirteet, sekä kehittää eettisiä ohjeistuksia mallien käyttöön. Tällaiset toimet ovat välttämättömiä, jotta voidaan varmistaa mallien tuottaman tiedon oikeudenmukaisuus ja luotettavuus (Bender ym. 2021).

Kokonaisuudessaan kielimallien koulutus ja koulutusdatan monimuotoisuus ovat olennaisia tekijöitä mallien tehokkuudessa ja oikeudenmukaisuudessa. Jatkuvasti kehittyvät menetel-

mät ja eettiset puitteet muovaavat tekoälyn tulevaisuutta, mikä avaa uusia mahdollisuuksia ja haasteita kielenkäsittelyn saralla.

4.4 Ymmärryksen rajallisuus ja haasteet

Vaikka kielimallit ovat kehittyneet huomattavasti viime vuosina, niiden ymmärryksen rajallisuus on edelleen merkittävä haaste. Mallit voivat simuloida ymmärrystä tuottamalla koherenttia ja kontekstuaalisesti sopivaa tekstiä, mutta ne eivät kykene ihmisen kaltaiseen abstraktiin ajatteluun, tunteiden ymmärtämiseen tai syvälliseen kontekstuaaliseen analyysiin. Linzen (2022) tuo esiin, että vaikka nykyaikaiset kielimallit pystyvät suoriutumaan monimutkaisista kieliopillisista tehtävistä, ne eivät vielä kykene täysin ymmärtämään kielen syvällisempiä merkityksiä ja monimutkaisia ihmisen kommunikaation muotoja (Linzen 2020).

Kielimallit ja ihmisaivot toimivat eri tavoin. Kielimallit voivat käsitellä paljon tietoa samanaikaisesti, mutta ihmisten keskittyminen on rajallisempaa ja valikoivampaa. Tämä ero vaikuttaa siihen, miten hyvin kielimallit pystyvät jäljittelemään ihmisen kielenkäsittelyä. Tutkimukset ovat myös paljastaneet, että kielenkäsittely ihmisaivoissa riippuu tietyistä kieliopillisista rakenteista. Tämän tiedon hyödyntäminen voi auttaa meitä kehittämään tehokkaampia kielimalleja (Kuribayashi ym. 2022).

5 Ymmärryksen määrittely tekoälyssä

Tässä osiossa syvennytään ”aitoon ymmärtämiseen” tekoälyssä ja siihen, kuinka tämä käsite eroaa ihmisen kognitiivisesta prosessista. Tarkastelun kohteena ovat Turingin testin ja Dreyfusin veljesten näkemykset tekoälyn rajoituksista ymmärryksessä. Keskitymme myös nykyaikaisten kielimallien kykyyn prosessoida kieltä, vertaillen tätä ihmisen kielelliseen ymmärrykseen ja pohtien, miten tilastollinen analyysi vaikuttaa näiden mallien toimintaan ja rajoituksiin.

5.1 Mitä on ”aito ymmärtäminen”?

”Aito ymmärtäminen” tekoälyn yhteydessä herättää laajaa keskustelua ja pohdintaa. Tämä käsite, joka ihmisille luontaisesti sisältää kyvyn reflektoida ja olla tietoinen oppimastaan, muodostuu tekoälyn kohdalla moniulotteisemmaksi haasteeksi. Alan Turingin esittelemä Turingin testi on klassinen esimerkki yrityksestä arvioida tekoälyn kykyä jäljitellä ihmisen älykkyyttä niin taidokkaasti, että ihmisen on vaikea erottaa sitä toisesta ihmisestä. Tämä testi avaa merkittäviä näkökulmia tekoälyn kykyihin, mutta se ei kuitenkaan ota kantaa koneiden todelliseen tietoisuuteen tai syvempään ymmärrykseen (Turing 1950). Täydentäen tätä ajatusta, Dreyfus ja Dreyfus (1987) ovat esittäneet näkemyksen, että aito ymmärtäminen sisältää elementtejä, kuten intuitiota ja kokemuspohjaista tietämystä, jotka ovat tekoälylle saavuttamattomia ominaisuuksia. Heidän mukaansa tekoäly, vaikka se kykenee prosessoimaan tietoa ja suorittamaan monimutkaisia tehtäviä, ei voi saavuttaa ihmismielen tasoista ymmärrystä ilman subjektiivista kokemusta ja tietoisuutta (Dreyfus ja Dreyfus 1987).

Tekoälyn tutkimuksessa on keskeistä ymmärtää ihmisen kognition moninaisuus. Teokset (Kahneman 2011) ja (Kahneman, Sibony ja Sunstein 2021), tarjoavat arvokkaita näkökulmia tähän monimuotoisuuteen. Kahneman erottaa ihmisen ajattelun kahteen järjestelmään: nopeaan, intuitiiviseen ajatteluun ja hitaaseen, loogiseen päättelyyn. Tämä jaottelu auttaa ymmärtämään tekoälyn rajoituksia ja mahdollisuuksia. Esimerkiksi neuroverkkomallit, jotka tuottavat nopeasti todennäköisiä vastauksia, vastaavat jossain määrin ihmisen intuitiivista ajattelua. Kuten ihmiset, myös nämä mallit voivat tuottaa virheellisiä tuloksia ja ovat alttiita

kognitiivisille vinoumille. Tämä korostaa, että tekoälyn kyky matkia ihmisen ajattelua ei aina merkitse samanlaista tiedon prosessointia tai ymmärrystä. Ihmisen kyky syvälliseen pohdintaan ja kokonaisvaltaiseen ymmärrykseen on monimuotoinen ja vaikeasti saavutettavissa oleva ominaisuus tekoälylle.

Lisäksi on tärkeää huomioida, että ihmisen kognitio ei ole täysin ymmärretty eikä helposti mallinnettavissa. Tämä monimutkaisuus luo haasteita tekoälyn kehittämiselle pyrittäessä saavuttamaan syvempi ja aito ymmärryksen taso.

5.2 Miten tekoäly ymmärtää kieltä?

Kun keskustelemme tekoälyn kyvystä ymmärtää kieltä, on ensiarvoisen tärkeää tunnistaa, että tässä yhteydessä käytetty ”ymmärtäminen” poikkeaa huomattavasti siitä, miten me ihmiset ymmärrämme kieltä. Syväoppimisen merkitys kielimallinnuksessa on ollut keskeinen aiemmissa keskusteluissani, mutta on tärkeää muistaa, että ennen syväoppimisen aikakautta tekoälyn kehityksessä olivat vahvasti läsnä semanttiset verkot ja symboliset lähestymistavat. Russell ja Norvig (2020) ovat nostaneet esille, kuinka nämä varhaisemmat mallit pyrkivät matkimaan inhimillisen kielen ymmärtämisen logiikkaa ja rakenteita, vaikka ne eivät päässeetkään käsiksi syväoppimisen myötä saavutettavaan monimutkaisuuden ja kielirakenteiden syvälliseen hahmottamiseen.

Lisäksi, vaikka nykyajan kielimallit kykenevät tuottamaan tekstiä, joka vaikuttaa vaikuttavalta ja vastaamaan moniin monimutkaisiin kysymyksiin, ne perustuvat pääasiassa tilastolliseen analyysiin ja eivät todellisuudessa ”ymmärrä” kieltä samalla tavoin kuin ihmiset. Nämä mallit ovat alttiita virheille ja voivat antaa harhaanjohtavia vastauksia, erityisesti kohdatessaan epätavallisia tai uusia tilanteita. Marcus (2018) on korostanut, että vaikka tekoälyn kyky kielenkäsittelyssä on edennyt pitkälle, sen ymmärrys kielen ja kontekstin suhteen on edelleen varsin rajallista. Tämä ymmärrys nojaa suurilta osin käytettävissä olevan datan määrään ja siihen, miten tämä data on tilastollisesti prosessoitu.

6 Yhteenveto

Tutkielmassa käsitellään suurten kielimallien, kuten GPT:n, kielen ymmärtämisen kykyä sekä niiden kielen tuottamisen ja ymmärtämisen mahdollistavia mekanismeja ja rakenteita. Tässä yhteenvedossa analysoidaan, onko näiden mallien kielen ymmärtäminen todellista vai ainoastaan vaikutelman luomista, ja tarkastellaan niiden keskeisiä toimintaperiaatteita.

Tutkielmassa tarkastellaan suurten kielimallien kielen tuottamisen ja ymmärtämisen keskeisiä mekanismeja ja rakenteita. Keskeisenä elementtinä tässä on syväoppimisen ja erityisesti transformer-arkkitehtuurin rooli. Transformer-mallien ”self-attention-mekanismi” korostuu, sillä se mahdollistaa olennaisten tekstielementtien erottelun ja syvällisen ymmärtämisen. Lisäksi ennakoivat mallit, kuten GPT-sarjan mallit, oppivat ennustamaan seuraavia sanoja tai lauseita laajoista tekstiaineistoista, syventäen ymmärrystämme kieliopillisista ja semanttisista rakenteista.

Lopuksi tutkielmassa käsitellään, missä määrin suurten kielimallien ymmärrys on vain vaikutelman luomista ja missä määrin sitä voidaan pitää ”aitona” ymmärtämisenä. Vaikka nämä mallit pystyvät tuottamaan koherenttia ja kontekstuaalisesti relevanttia tekstiä, niiden kyvyt eivät ole verrattavissa ihmismielen syvälliseen kielen ymmärtämiseen ja abstraktiin pohdintaan. Mallien tuottama ”ymmärrys” perustuu pääasiassa datan tilastolliseen analyysiin ja prosessointiin, eikä niin paljon syvälliseen kontekstuaaliseen analyysiin. Tämän vuoksi, vaikka suuret kielimallit kykenevät luomaan vakuuttavan vaikutelman kielen ymmärtämisestä, niiden todellinen ymmärryskyky on rajoitettu. Ne eivät kykene ihmisen kaltaiseen abstraktiin ajatteluun tai tunteiden ymmärtämiseen. Vaikka suuret kielimallit ovat kehittyneet tuottamaan tekstiä, joka antaa vaikutuksen ymmärtämisestä, niiden kyky ymmärtää kieltä ”aidosti” on edelleen rajoittunutta.

Lähteet

Aniba, Mohamed Radhouene, Sophie Siguenza, Anne Friedrich, Frédéric Plewniak, Olivier Poch, Aron Marchler-Bauer ja Julie Dawn Thompson. 2008. “Knowledge-based expert systems and a proof-of-concept case study for multiple sequence alignment construction and analysis”. *Briefings in Bioinformatics* 10 (1): 11–29. https://doi.org/10.1093/bib/bbn045. https://doi.org/10.1093/bib/bbn045.

Bellman, R. 1978. *An Introduction to Artificial Intelligence: Can Computers Think?* Boyd & Fraser Pub. Co. ISBN: 0878350667. https://searchworks.stanford.edu/view/2762753.

Bender, Emily M, Timnit Gebru, Angelina McMillan-Major ja Shmargaret Shmitchell. 2021. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency,* 610–623. https://doi.org/10.1145/3442188.3445922. https://doi.org/10.1145/3442188.3445922.

Bengio, Yoshua ja Christian Jauvin Rejan Ducharme Pascal Vincent. 2003. “A neural probabilistic language model”. *Journal of machine learning research* 3 (Feb): 1137–1155. https://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf.

Bengio, Yoshua, Patrice Simard ja Paolo Frasconi. 1994. “Learning long-term dependencies with gradient descent is difficult”. *IEEE Transactions on Neural Networks* 5 (2): 157–166. https://doi.org/10.1109/72.279181. https://ieeexplore.ieee.org/document/279181.

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein ym. 2021. *On the Opportunities and Risks of Foundation Models.* https://doi.org/arXiv:2108.07258. https://arxiv.org/abs/2108.07258.

Brown, Tom B, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell ym. 2020. “Language Models are Few-Shot Learners”. *Advances in Neural Information Processing Systems* 33:1877–1901. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Charniak, E. ja D. McDermott. 1985. “Introduction to artificial intelligence”, https://www.osti.gov/biblio/7249333.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee ja Kristina Toutanova. 2018. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. *arXiv preprint arXiv:1810.04805,* https://doi.org/10.48550/arXiv.1810.04805. https://doi.org/10.48550/arXiv.1810.04805.

Dreyfus, Hubert L. ja Stuart E. Dreyfus. 1987. *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer.* 2:110–111. 2. IEEE. https://doi.org/10.1109/MEX.1987.4307079. https://ieeexplore.ieee.org/document/4307079.

Goldberg, Yoav. 2017. *Neural Network Methods for Natural Language Processing.* CCXCII, 20. Synthesis Lectures on Human Language Technologies. Springer Cham. ISBN: 978-3-031-01037-8. https://doi.org/https://doi.org/10.1007/978-3-031-02165-7.

Goodfellow, Ian, Yoshua Bengio ja Aaron Courville. 2016. *Deep Learning.* Http://www.deeplearningbook.org. MIT Press.

Haugeland, J. 1985. *John Haugeland. Artificial intelligence: the very idea. Bradford books. The MIT Press, Cambridge, Mass., and London, 1985, ix 289 pp.* The MIT Press, Cambridge University Press. https://doi.org/10.2307/2274541. https://www.cambridge.org/core/journals/journal-of-symbolic-logic/article/abs/john-haugeland-artificial-intelligence-the-very-idea-bradford-books-the-mit-press-cambridge-mass-and-london-1985-ix-289-pp/81C8ECEE45C224FB659B4637809DCE6D.

Hirschberg, Julia ja Christopher D. Manning. 2015. “Advances in natural language processing”. *Science* 349 (6245): 261–266. https://doi.org/10.1126/science.aaa8685. https://www.science.org/doi/abs/10.1126/science.aaa8685.

Hovy, Dirk ja Shannon L. Spruit. 2016. “The social impact of natural language processing”. Teoksessa *Proceedings of the 54th annual meeting of the association for computational linguistics (Volume 2: Short Papers),* 591–598. Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-2096. https://aclanthology.org/P16-2096.

Jurafsky, Daniel ja James H. Martin. 2019. *Speech and Language Processing.* Pearson. https://web.stanford.edu/~jurafsky/slp3/.

Kahneman, Daniel. 2011. *Thinking, Fast and Slow.* Farrar, Strauss / Giroux. ISBN: 9780374275631. https://search.worldcat.org/title/815190472?oclcNum=815190472.

Kahneman, Daniel, Olivier Sibony ja Cass R. Sunstein. 2021. *Noise: A Flaw in Human Judgment.* Little, Brown Spark. ISBN: 9780316451406. https://search.worldcat.org/title/1288124022.

Kapitanova, Krasimira ja Sang H Son. 2012. "Machine Learning Basics". Teoksessa *Intelligent Sensor Networks: The Integration of Sensor Networks, Signal Processing and Machine Learning,* 3–29. CRC Press. ISBN: 9781439892824. https://www.taylorfrancis.com/books/oa-edit/10.1201/b14300/intelligent-sensor-networks-fei-hu-qi-hao.

Kuribayashi, Tatsuki, Yohei Oseki, Ana Brassard ja Kentaro Inui. 2022. "Context Limitations Make Neural Language Models More Human-Like". *ar5iv,* https://arxiv.org/abs/2205.11463.

Kurzweil, R. 1990. *The Age of Intelligent Machines.* Association of American Publishers, Inc. ISBN: 9780262111218. https://mitpress.mit.edu/9780262111218/the-age-of-intelligent-machines/.

LeCun, Yann, Yoshua Bengio ja Geoffrey Hinton. 2015. "Deep learning". *Nature* 521 (7553): 436–444. https://doi.org/10.1038/nature14539. https://www.cs.toronto.edu/~hinton/absps/NatureDeepReview.pdf.

Linzen, Tal. 2020. "How Can We Accelerate Progress Towards Human-like Linguistic Generalization?", https://doi.org/arXiv:2005.00955. https://arxiv.org/abs/2005.00955.

Marcus, Gary. 2018. "Deep learning: A critical appraisal". *arXiv preprint arXiv:1801.00631,* https://doi.org/arXiv:1801.00631. https://arxiv.org/abs/1801.00631.

Mitchell, Tom M. 1997. *Machine Learning.* 414. Http://www.cs.cmu.edu/~tom/mlbook.html. McGraw-Hill Education. ISBN: 0070428077, 9780070428072, 0071154671, 9780071154673.

Murphy, Kevin P. 2012. *Machine Learning: A Probabilistic Perspective.* 1104. Cambridge, MA, USA: The MIT Press. ISBN: 9780262018029. https://doi.org/10.5555/2380985. https://mitpress.mit.edu/9780262018029/machine-learning/.

Naveed, Humza, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes ja Ajmal Mian. 2023. "A Comprehensive Overview of Large Language Models", https://arxiv.org/abs/2307.06435.

Nilsson, N. J. 1998. *Artificial Intelligence : a new synthesis.* Morgan Kaufmann Publishers. ISBN: 9781558605350, 9781558604674, 9780080499451. https://search.worldcat.org/title/1148016147.

Poole, D., Alan K. Mackworth ja Randy Goebel. 1998. *Computational intelligence : a logical approach.* Oxford University Press. ISBN: 9780195102703. https://global.oup.com/ushe/product/computational-intelligence-9780195102703?cc=fi&lang=en&.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei ja Ilya Sutskever. 2019. *Language Models are Unsupervised Multitask Learners.* https://openai.com/blog/better-language-models/.

Rich, E. ja K. Knight. 1991. *Artificial intelligence.* McGraw-Hill. ISBN: 9780070522633, 9780071008945, 9780074600818. https://search.worldcat.org/title/22593030.

Rumelhart, David E., Geoffrey E. Hinton ja Ronald J. Williams. 1986. "Learning representations by back-propagating errors". *Nature* 323:533–536. https://api.semanticscholar.org/CorpusID:205001834.

Russell, Stuart J ja Peter Norvig. 2020. *Artificial Intelligence: A Modern Approach.* 1132. Prentice Hall series in artificial intelligence. Http://aima.cs.berkeley.edu. Prentice Hall. ISBN: 9780136042594, 0136042597, 9780132071482, 0132071487.

Turing, Alan M. 1950. "Computing Machinery and Intelligence". *Mind* 59 (236): 433–460. https://doi.org/10.1093/mind/LIX.236.433. https://doi.org/10.1093/mind/LIX.236.433.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser ja Illia Polosukhin. 2017. “Attention is all you need”. Teoksessa *Advances in neural information processing systems,* 30:5998–6008. Curran Associates, Inc. ISBN: 9781510860964. https://doi.org/10.5555/3295222.3295349. https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

Weidinger, Laura, John Mellor, Matt Riedl, Debapriya Chakraborty, Emily M Bender ja Ryan Cotterell. 2021. “Ethical and Social Risks of Harm from Language Models”. *arXiv preprint arXiv:2112.04359,* https://doi.org/10.48550/arXiv.2112.04359. https://arxiv.org/abs/2112.04359.

Winston, P. H. 1992. *Artificial Intelligence.* Addison-Wesley Publishing Company. ISBN: 0201533774, 9780201533774. https://books.google.fi/books?id=b4owngEACAAJ.

Xie, Sang Michael, Shibani Santurkar, Tengyu Ma ja Percy Liang. 2023. “Data Selection for Language Models via Importance Resampling”, https://arxiv.org/abs/2302.03169.