TURUN YLIOPISTO

Tiedon eristäminen materiaalitieteiden teksteistä suurilla kielimalleilla

Materiaalitekniikan
Kanditutkielma

Laatija:
Viljami Nuutinen

19.6.2024
Turku

Kanditutkielma

Oppiaine: Materiaalitekniikka
Tekijä: Viljami Nuutinen
Otsikko: Tiedon eristäminen materiaalitieteiden teksteistä suurilla kielimalleilla
Ohjaaja: FM Matilda Sipilä
Sivumäärä: 23 sivua
Päivämäärä: 19.6.2024

Materiaali-informatiikka on materiaalitieteen haara, jossa hyödynnetään laskennallisia metodeja materiaalien ja niiden ominaisuuksien tutkimiseen ja kehittämiseen. Materiaali-informatiikan tutkimus vaatii kuitenkin suuria määriä dataa, jonka saatavuus on haastavaa johtuen materiaalitieteiden tietokantojen vajavaisuudesta. Materiaalitieteen julkaisujen määrä kasvaa jatkuvasti, mutta alan monimuotoisuuden vuoksi tiedon esittämistavat ovat vaihtelevia ja halutun tiedon löytäminen työlästä.

Suurilla kielimalleilla (kuten BERT (Bidirectional Encoder Representations from Transformers) tai GPT (Generative Pre-trained Transformer)), voidaan analysoida suuria määriä tekstiä automaattisesti ja eristää niistä arvokasta dataa materiaaleista, josta voidaan luoda tietokantoja hyödynnettäväksi materiaali-informatiikan sovelluksissa. Tiedon eristäminen materiaalitieteiden teksteistä on haastavaa, koska erilaisia tutkittavia materiaaliluokkia on paljon ja merkintätavat vaihtelevat alalla. Teksti on myös vaikea datan laji, koska se voi esiintyä erilaisissa muotoiluissa, mikä mutkistaa sen käsittelyä.

Tutkielmassa tarjotaan katsaus kieliteknologioiden käyttöön materiaalitieteissä, aiheeseen liittyvään termistöön ja materiaalitekniikan käyttöön kehitettyihin kielimalleihin. Tutkielma keskittyy käsittelemään tekstimuotoista tiedon eristämistä BERT-kielimalleilla. Suurilla kielimalleilla tiedon eristäminen on materiaalitieteissä alkutekijöissään ja siihen liittyviä haasteita on paljon. Kieliteknologioiden kehittyessä tiedon eristäminen suurilla kielimalleilla on vaikeuksista huolimatta lupaava työkalu tulevaisuuden materiaalien kehityksessä.

Avainsanat: tiedon eristäminen, materiaali-informatiikka, suuret kielimallit, kieliteknologia

Sisällysluettelo

1 Johdanto

Materiaalitekniikan tai materiaalitieteen pyrkimys on ymmärtää ja mallintaa materiaalien ominaisuuksia erilaisiin sovelluksiin edistääkseen ihmiskunnan kehitystä. Materiaalitiede on laaja tieteenala, joka kattaa osa-alueita kaikista luonnontieteistä ja kaikki nyky-yhteiskunnan tuotteet hyödyntävät materiaaleja. Tietoa materiaaleista on paljon eri muodoissa, joista yksi on tekstimuodoissa olevat tieteelliset julkaisut.

Digitalisaatio sekä tekoälyn ja koneoppimisen kehittyminen ovat lisänneet saatavilla olevan tiedon määrää ja mahdollistaneet erilaisia dataa hyödyntäviä lähestymistapoja tutkimukseen. Luonnollisen kielen käsittely suurilla kielimalleilla on tehnyt mahdolliseksi valtavien tekstimassojen koneellisen tulkitsemisen ja käsittelyn erilaisia sovelluskohteita varten. Tämä jatkuva kehitys enteilee tiedon saatavuuden paranemista lukuisilla eri aloilla ja uusien edistyksellisten tutkimusmenetelmien syntyä.

Materiaali-informatiikka hyödyntää laajoja tietoaineistoja ja koneoppimista uusien materiaalien tehokkaampaan kehittämiseen. Nämä menetelmät vaativat kuitenkin suuria määriä soveltuvaa dataa. Valtavista määristä tekstiä, jota on materiaalitieteen julkaisuissa, voidaan tiedon eristämisen avulla kerätä arvokasta tietoa dataksi materiaali-informatiikan käyttötarkoituksiin. Suurten kielimallien muokkaaminen materiaalitieteiden alalle soveltuviksi tarjoaa lupaavaa ratkaisua monimuotoisen tiedon eristämiseksi suurista määristä erilaisia tekstimuotoisia julkaisuja.

Tämän tutkielman tarkoitus on esitellä tiedon eristämistä materiaalitieteiden teksteistä suurilla kielimalleilla kirjallisuuskatsauksena ja vastata seuraaviin tutkimuskysymyksiin:

1. Miksi tietoa eristetään materiaalitieteiden teksteistä?
2. Millaista tietoa teksteistä eristetään?
3. Miten tiedon eristäminen tehdään?

Tutkielmassa käydään ensin läpi luonnollisen kielen käsittelyn termistöä ja menetelmiä. Tämän jälkeen keskitytään kieliteknologioiden käyttöön materiaalitieteiden näkökulmasta. Viimeisessä osioissa esitellään erilaisia materiaalitekniikan kielimalleja, perehdytään tarkemmin kolmeen sekä vertaillaan niiden koulutusta ja toimintaa. Lopuksi pohditaan vielä materiaalitekniikan kielimallien haasteita ja mahdollisuuksia. Kaikki tutkielmassa käytetyt kuvat ovat tutkielman laatijan tekemiä.

2 Luonnollisen kielen käsittely

Luonnollisen kielen käsittely (engl. natural language processing, NLP) on tieteenala, joka toimii kielitieteen ja tietojenkäsittelytieteen leikkauskohdassa. Luonnollisen kielen käsittelyn tarkoituksena on erilaisten ihmisten tuottamien luonnollisen kielen tekstien tai äänitteiden, kuten kirjojen, tieteellisten artikkelien tai sähköpostien käsittely sekä ymmärtäminen tietokoneiden avulla. NLP mahdollistaa myös tietokoneavusteisen ihmiskielen tuottamisen.[1] Luonnollisen kielen käsittely tarjoaa uusia laskennallisia menetelmiä kielten tunnistamiseen, kääntämiseen, tuottamiseen, tiedon eristämiseen sekä muihin sovelluksiin, joita on esitetty enemmän kuvassa 1.[2] Tässä tutkielmassa keskitytään pääosin kirjoitetun tiedon eristämisen menetelmiin sekä sovelluskohteisiin.

Kuva 1. Luonnollisen kielen käsittelyn sovelluksia.

Nykymallinen luonnollisen kielen käsittely hyödyntää laajalti eri koneoppimisen keinoja ja tekee mahdolliseksi monimutkaisten tekstien käsittelyn ennustettavilla ja toistettavilla tuloksilla.[2] Koneoppiminen on tekoälyn osa-alue, jonka tarkoituksena on kehittää erilaisia keinoja, joilla tietokonejärjestelmät kykenevät käsittelemään ja soveltamaan tietoa.[3] Koneoppiminen mahdollistaa aiemmin opitun avulla yhteyksien muodostamisen erilaisten

muuttujien välillä ja suurien datamäärien käsittelyn.[4] Yleisimmin käytetyt koneoppimismallit ovat ohjattu oppiminen ja ohjaamaton oppiminen. Ohjatussa oppimisessa konetta opetetaan määritellyn datan avulla tunnistamaan syötteitä ja luokittelemaan tietoa halutulla tavalla. Datan luokitteleminen ja koneiden opettaminen ovat kuitenkin työläitä ja hitaita prosesseja. Ohjaamattomassa oppimisessa koneelle syötetään luokittelematonta dataa, josta kone pyrkii löytämään samankaltaisuuksia ja luokittelemaan dataa itsenäisesti.[3]

Kone tai ohjelma täytyy opettaa ymmärtämään käsiteltävän kielen perussäännöt ja kielioppi, jotta sen on mahdollista lukea, ymmärtää, prosessoida ja tuottaa tekstiä. Luonnollisen kielen käsittelyssä tämän oppimisprosessin erilaisia vaiheita ovat mm. lauseiden erittely (engl. sentence segmentation), tokenointi (engl. tokenization), typistäminen (engl. stemming) ja perusmuotoistaminen (engl. lemmatization), sanaluokkajäsennys (engl. part-of-speech tagging) ja nimettyjen entiteettien tunnistus (engl. named entity recognition, NER).[5]

Pidemmän tekstikokonaisuuden käsittely aloitetaan jakamalla teksti yksittäisiin lauseisiin, joiden käsittelyä on helpompi jatkaa. Lauseiden erittely voidaan toteuttaa useimmiten jakamalla lauseet välimerkkien kohdalta. Käsiteltävät tekstit eivät tosin aina ole oikein muotoiltuja, jolloin voidaan turvautua tekstin erittelyyn käyttäen apuna sekvenssioppimista, jossa kieliopin ja asiayhteyksien pohjalta lauseet eritellään oikealla tavalla.[5]

Tokenoinnissa lauseet jaetaan tokeneihin, jotka voivat olla sanoja, saman sanan eri muotoja tai osia sanoista. Tokenointi voidaan toteuttaa yksinkertaisimmillaan jakamalla jokainen välillä erotettu sana omaksi tokenikseen. Tokenit voidaan joskus muuttaa yksinkertaisempaan muotoon vaihtamalla kaikki kirjaimet pieniksi sekä poistamalla erikoismerkit. Yksittäiset kirjaimet, kuten englannin artikkeli ’a’ tai lausekkeet, kuten ’machine learning’ voivat olla myös tokeneita.[6]

Typistäminen on prosessi, jossa tokeneita paloitellaan niiden yksinkertaisimpaan muotoon poistamalla sanaliitteet. Esimerkiksi sanat ’kone’ ja ’koneissa’ typistyisivät molemmat muotoon ’kone’. Näin jokaiselle sanan eri muodolle ei tarvitse olla omaa tokenia, mikä tehostaa ohjelman käsittelynopeutta. Kaikkia tokeneita ei pystytä kuitenkaan typistämään, koska silloin sanan merkitys voi muuttua. Perusmuotoistamisessa tokenit muutetaan sanakirjamääritelmiä avuksi käyttäen perusmuotoon. Typistämällä sanat ’kuu’ ja ’kuulla’ saisivat saman vartalon ’kuu’, ja jälkimmäinen sana menettäisi alkuperäisen merkityksensä. Perusmuotoistamalla sanojen merkitys säilyy, koska sanoja käsittelevä algoritmi on laadittu löytämään sanan merkitystä vastaava perusmuoto.[5]

Sanaluokkajäsennyksen avulla pystytään välttämään tilanteet, joissa samalla sanalla on useita merkityksiä eri yhteyksissä. Tietyt sanat voivat kuulua useampiin sanaluokkiin. Esimerkiksi sana ’veto’ voi olla verbi tai substantiivi. Hyödyntämällä korpuksia, eli määriteltyjä kokoelmia valtavista määristä kirjoitetun kielen tekstejä, pystytään löytämään sanoille yhteyksiä tietynlaisissa lauseissa ja konteksteissa. Näin onnistutaan parhaimmillaan määrittämään sanoille oikea merkitys 96 %:n tarkkuudella.[5,7]

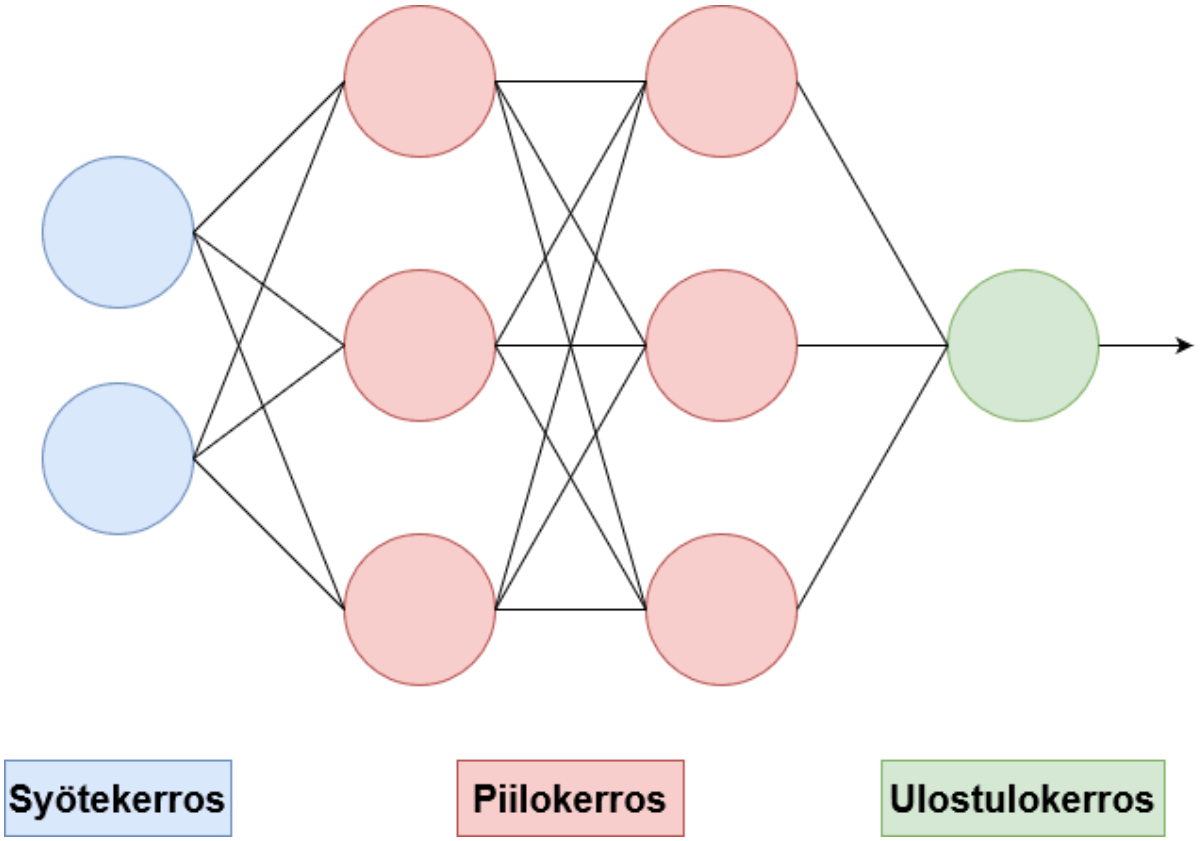
Nimettyjen entiteettien tunnistuksen (NER) avulla tekstistä kyetään tunnistamaan tiettyjä sanoja, sanojen yhdistelmiä tai muita merkkejä ja kategorisoimaan niitä. Entiteettejä voidaan rajata ja luokitella millä vain tavoin tarvittavien parametrien mukaan. Entiteetti voi olla esim. materiaali, materiaalin ominaisuus tai sovellus.[8] Kuvassa 2 on esimerkki tekstistä tunnistetuista entiteeteistä. NER on tärkeässä osassa luonnollisen kielen käsittelyä varsinkin eristäessä tietyn tyyppistä tietoa, kuten materiaalien ominaisuuksia mittavista tekstuaalisista tietokannoista.

Kenties merkittävin magnesiumin ominaisuuksista on sen tiheys 1,7g/cm³
Tunnisteet: Materiaali, Ominaisuus, Luku, Yksikkö

Kuva 2. Nimettyjen entiteettien tunnistaminen materiaalitieteen tekstistä.[9]

2.1 Transformer-neuroverkot

Neuroverkot ovat koneoppimismalleja, jotka mukailevat ihmisaivojen tiedonkäsittelyprosessia. Neuroverkot koostuvat useista neuroneista, eli prosessiyksiköistä, jotka vastaanottavat ja käsittelevät tietoa sekä tuottavat siitä päätelmiä. Neuronien kyky muuttaa syötettään, ja näin muokata yhteyksiä neuronien välillä tekee niistä tehokkaita työkaluja koneoppimisessa.[10] Neuroverkoissa neuroneja voidaan jakaa eri kerroksiin niiden tehtävän mukaan. Kuvassa 3 on esitetty yksinkertainen neuroverkkorakenne, jossa käsiteltävä tieto otetaan vastaan syötekerroksessa, käsitellään piilokerroksissa ja syötetään lopulta käsiteltynä tuloksena ulostulokerroksesta. Syväoppimisella viitataan koneoppimismenetelmiin, joissa hyödynnetään useita piilokerroksia sisältäviä neuroverkkoja.[1]



Kuva 3. Mallinnus neuroverkosta, jossa on syötekerros, kaksi piilokerrosta ja ulostulokerros.

Transformer-neuroverkkoarkkitehtuuri on vuonna 2017 esitelty syväoppimismalli, jota hyödynnetään etenkin NLP-tehtävissä. Transformer-arkkitehtuuri eroaa merkittävästi aiemmista malleista sen hyödyntämien huomiomekanismien (engl. attention mechanism) ansiosta. Huomiomekanismit kiinnittävät huomiota käsiteltävän syötteen oleellisiin ominaisuuksiin ja pystyvät oppimaan konteksteja. Huomiota pystytään kiinnittämään syötteen kaikkien tunnisteiden välillä samanaikaisesti. Tämä mahdollistaa suurten tekstimäärien käsittelemisen nopeasti.[11]

2.2 Suuret kielimallit

Suuret kielimallit (eng. large language models) ovat esiopetettuja koneoppimisohjelmia, jotka kykenevät tuottamaan ihmiskielimäistä tekstiä muodostamalla yhteyksiä eri sanojen välillä ja ennustamalla kontekstiriippuvaisia jatkumoita sanoille ja lauseille. Nykyisin lähes kaikki suuret kielimallit perustuvat Transformer-arkkitehtuuriin ja kykenevät esimerkiksi tuottamaan vastauksia kysymyksiin sekä tiivistämään tietoa esiopetetun datan perusteella. Suuret kielimallit koulutetaan laajoilla kieliaineistoilla, minkä johdosta ne oppivat luomaan vahvasti kontekstiin pohjautuvia syötteitä.[12]

GPT-kielimallit (Generative Pre-trained Transformer) ovat yleiseen käyttöön soveltuvia kielimalleja, jotka tuottavat syötteitä tekemällä ennusteita aiemmin koulutettuun valtavaan datamäärään perustuen. GPT-mallien vahvuus on tekstin tuottaminen ja niitä voidaan hyödyntää useissa eri tehtävissä, kuten koodin kirjoittamisessa sekä kielten kääntämisessä. GPT-kielimallit ovat suunnannäyttäjiä tulevaisuuden keskustelubottien ja virtuaaliavustajien kehityksessä. Vuonna 2020 julkaistun GPT-3:n kouluttamisessa käytettiin lähes 500 miljoonaa tokenia, jotka koostuivat nettisivujen teksteistä, Wikipedia-artikkeleista ja kirjallisuuskorpuksista.[13]

BERT-kielimallit (Bidirectional Encoder Representations from Transformers) ovat Transformer-arkkitehtuuria käyttäviä kielimalleja, jotka eivät GPT-mallien tavoin tuota uutta tekstiä, vaan analysoivat jo olemassaolevaa tekstiä. Ne ovat kaksisuuntaisia eli käsittelevät sanoja verraten niitä kaikkiin syötteessä aiemmin ja jälkeen tuleviin sanoihin. Näin saadaan kattavampi konteksti ja mahdollisimman tarkka luokittelu sekä lopputulos. BERT-kielimalleja hyödynnetään paljon luonnollisen kielen käsittelyn tehtävissä, kuten tekstin luokittelussa, hakutuloksien täsmentämisessä ja tiedon eristämisessä. BERT koulutettiin 3,3 miljardilla sanalla, jotka koostuivat englanninkielisestä Wikipediasta ja kirjoista koostuvasta korpuksesta.[14] Tässä tutkielmassa keskitytään pääasiassa erilaisten BERT-kielimallien eroihin ja mahdollisuuksiin materiaalitieteessä.

3 Kieliteknologian käyttö materiaalitieteessä

Teknologinen kehitys laskennallisissa työkaluissa, tehokkaammissa tutkimusmenetelmissä ja tiedon avoimemmassa sekä maailmanlaajuisessa jakamisessa on kasvattanut saatavilla olevan tieteellisen tiedon määrää olennaisesti. Datalähtöisessä materiaalitutkimuksessa tätä suurta määrää jo olemassa olevaa tietoa hyödynnetään koneoppimisen menetelmillä tutkimuksen tehostamiseksi. Koneoppimisen keinot mahdollistavat datan käsittelyn nopeammin, suuremmissa määrissä sekä monimutkaisempien ongelmien ratkaisun. Koneoppimismalleja on pystytty käyttämään tehokkaasti useissa sovelluksissa, kuten materiaalien ominaisuuksien ennustamisessa ja kokonaan uusien materiaalien havaitsemisessa monilla eri materiaalitieteen osa-alueilla.[15]

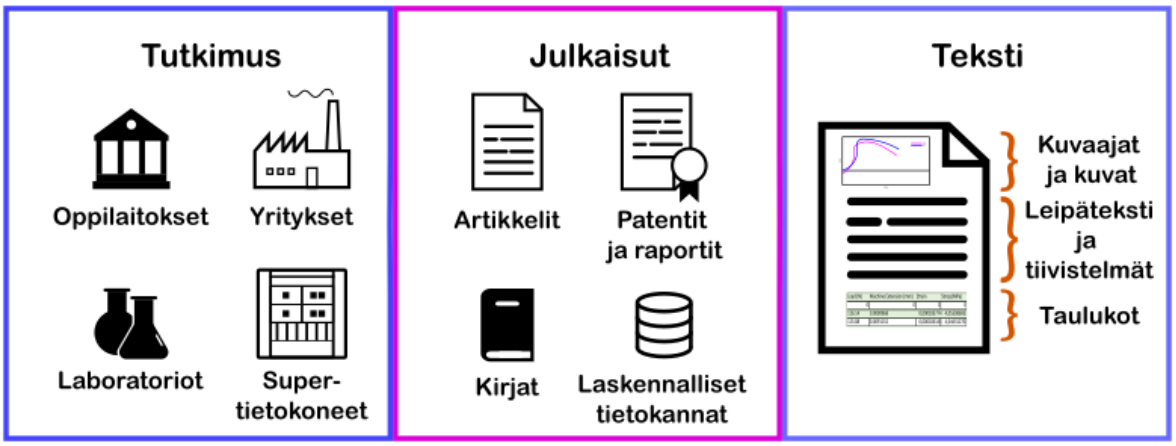
Datalähtöinen materiaalitutkimus vaatii kuitenkin datan olevan muodossa, jossa tietokoneohjelmat voivat hyödyntää sitä. Tällainen käyttökelpoinen ja empiiriseen tutkimukseen perustuva data on vielä rajallista, vaikka materiaalitieteiden kirjallisuutta onkin paljon. Materiaalitieteiden data on myös epäyhtenäistä, johtuen laajasta kirjosta erilaisia tutkittavia materiaalien luokkia, ominaisuuksia ja sovelluksia. Dataa kerätään myös vaihtelevin keinoin suurista tutkimuslaitoksista pieniin opetuslaboratorioihin sekä eri mittaluokissa atomitason kiderakenteista valtaviin laivojen rungon osiin.[16]

Koneoppimiseen soveltuvat materiaalitieteiden tietokannat ja tietoaineistot ovat puutteellisia, koska datan epäyhtenäisyys tekee niiden kokoamisesta työlästä ja aikaa vievää.[16,17] Materiaalitieteiden ensisijainen tiedonjakamisen keino ovat erilaiset julkaisut, jotka kattavat tieteellisten aikakausijulkaisujen artikkelit, patentit tai muut yritysten raportit ja selvitykset. Näistä erilaisista tieteellisistä julkaisuista voidaan kerätä ja eristää dataa suurilla kielimalleilla luonnollisen kielen käsittelyn keinoin. Eristetystä datasta voidaan luoda automaattisesti tietokantoja, joita voidaan hyödyntää muissa materiaali-informatiikan sovelluksissa tai perinteisessä tutkimuksessa.[16,18]

Kieliteknologia tehostaa tutkimusta materiaalitieteessä mahdollistamalla tutkimuksen tekemisen aiempia havaintoja kätevämmin hyödyntämällä. Suurista tekstimassoista voidaan tiivistää tarpeellinen informaatio materiaalitietokantoihin tietoa eristämällä. Tästä datasta voidaan havaita tiettyjä yhteyksiä ja kehityssuuntia, joiden avulla saadaan kartoitettua mahdollisia materiaaleja. Kattavat ja helppokäyttöiset tietokannat nopeuttavat

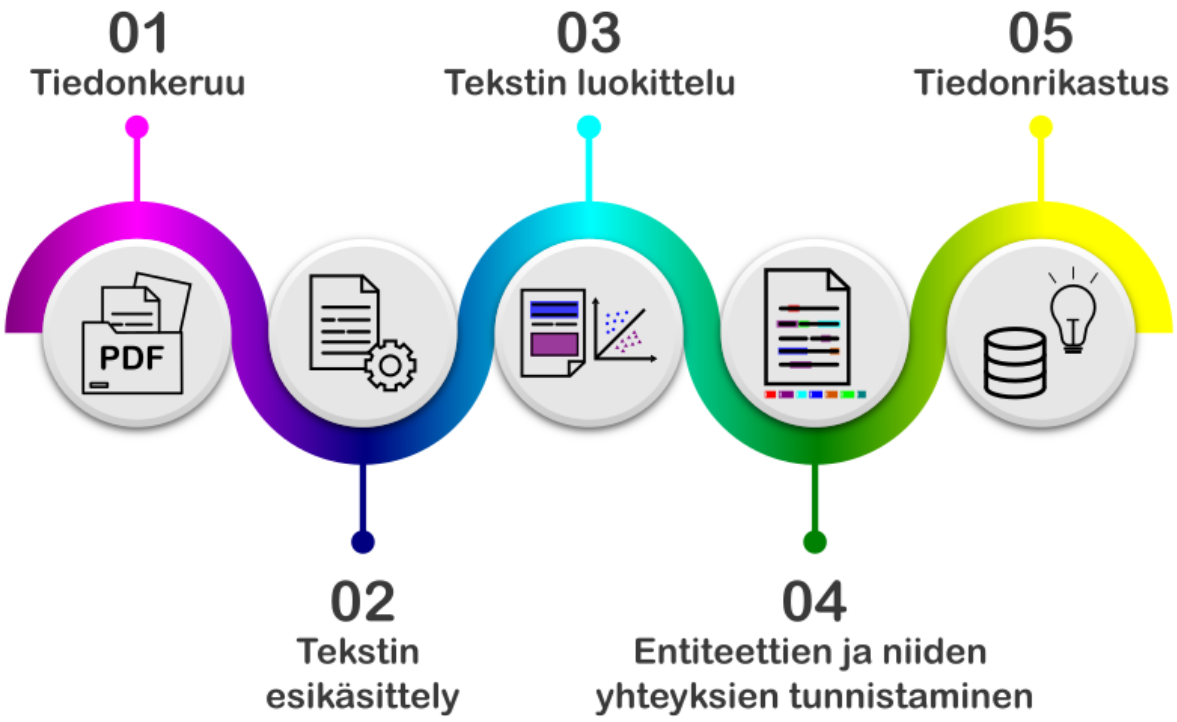
taustatutkimuksen tekemistä, vahvistavat tiedonvaihtoa eri tieteenhaarojen välillä ja voivat tehostaa erilaisia koneoppimismenetelmiä materiaalitieteen alalla.[16]

Luonnollisen kielen teksti on kuitenkin haastava datan laji. Erilaiset datan kirjaustavat vaihtelevat paljon ja yhdessä artikkelissa toivottu data tietystä materiaalista voi olla jakautunut useisiin eri muotoisiin tekstielementteihin. Tiedon eristämisen tavat voivat vaihdella leipätekstin, taulukoiden ja kuvien välillä. Kuvassa 4 on tiivistetty erilaisia materiaalitieteiden datanlähteitä ja esitystapoja. Erilaisten julkaisujen tiedostomuodot tuottavat myös haasteita, koska koneellinen tekstinkäsittely suurissa volyymeissa onnistuu parhaiten käsittelemättömän selväkielisen tekstin (engl. plain text) muodossa. Suurin osa julkaisuista on saatavilla merkintäkielimuodossa (engl. markup language), kuten HTML tai XML. Merkintäkieli on verrattain helppoa muuttaa selväkieliseksi. Vanhemmat PDF-muotoiset tai kuvana olevat tekstit ovat vaikeita käsitellä suuressa skaalassa automaattisesti.[16]



Kuva 4. Materiaalitieteiden erilaiset datanlähteet ja esitystavat.

Suurilla kielimalleilla tiedon eristäminen materiaalitieteissä toteutetaan yleisesti samanlaista kaavaa seuraavalla luonnollisen kielen käsittelyn työnkululla, jota mallinnetaan kuvassa 5. Työnkulun vaiheet ja järjestys voivat vaihdella riippuen eristettävän tiedon tyypistä ja käytetyistä työkaluista. Eri menetelmät voidaan luokitella useimmiten kuvassa 5 esitettyihin vaiheisiin, joita ovat tiedonkeruu, tekstin esikäsittely, tekstien luokittelu, entiteettien ja niiden yhteyksien tunnistus ja tiedonrikastus eli eristetyn tiedon analysointi ja hyödyntäminen materiaali-informatiikan sovelluksissa. Työnkulku mukailee vahvasti perinteistä materiaalitieteen tutkimusprosessia, jossa haetaan halutusta aiheesta tutkimusartikkeleita, valitaan aihetta vastaavat artikkelit, luetaan teksti ja kerätään sieltä haettu tieto ryhmiteltyyn muotoon.[16,17]



Kuva 5. Tiedon eristämisen työnkulku materiaalitieteissä.

Tiedonkeruun vaiheessa materiaalitieteiden tietoa voidaan tehokkaimmin eristää luomalla ensin suuria tietoaineistoja. Tietoaineistot kootaan hakemalla tieteellisiä artikkeleita halutusta materiaalitieteen aiheesta. Artikkeleita kootaan useista eri julkaisuista ja valikoidaan parhaiten soveltuvat. Tämä toteutetaan käyttämällä aiheeseen liittyviä hakusanoja ja lataamalla artikkeleita tieteellisten julkaisujen tietokannoista tai tekemällä verkon kaavintaa. Tiedonkeruussa voidaan myös hyödyntää patentteja, joissa data on selkeästi määritellyssä muodossa. Patentteja on tämän järjestelmällisen esitystyylin ansiosta helppo käsitellä. Patenttien saatavuus on kuitenkin rajallista ja niissä esitetty tieto voi olla epäsuoraa, koska patenttien tekijät eivät halua paljastaa kaikkea tietoaan.[16]

Tietoaineistojen kokoamisen jälkeen tekstit täytyy esikäsitellä, jotta niistä voidaan eristää haluttua tietoa. Esikäsittelyssä voidaan hyödyntää NLP-vaiheita, kuten tokenointia ja sanaluokkajäsennystä. Materiaalitieteiden tekstien tokenoinnissa täytyy huomioida alalle tyypilliset merkintätavat. Kemiallisten kaavojen ja materiaalien nimeämistavoissa pilkkujen, pisteiden, kaksoispisteiden ja väliviivojen käyttö vaatii erityistä tarkkuutta tokenoitaessa. Tieteellisissä artikkeleissa lauseopillinen rakenne on usein erilainen kuin yleiskäyttöisissä teksteissä. Kielimalleja koulutettaessa materiaalitieteiden kielen käsittelyyn on sovitettava ne

ymmärtämään tieteelliselle tekstille tyypillistä lauseoppia, kuten laajamittaista passiivin sekä imperfektin käyttöä ja pronominien puutetta.[16]

Tekstin luokittelun avulla datan järjestely ja hyödyntäminen helpottuvat. Artikkelin paloittelu erityyppisiin osiin, kuten leipätekstiin, taulukoihin ja kuviin nopeuttaa datan käsittelyä ja parantaa eristetyn tiedon yhtenäisyyttä. Erilaisten tekstielementtien käsittely erikseen mahdollistaa erilaisten prosessien hyödyntämisen tekstielementille sopivalla menetelmällä.

Tärkein vaihe tiedon eristämisessä materiaalitieteiden näkökulmasta on nimettyjen entiteettien tunnistaminen. Entiteettejä kartoittamalla pystytään luokittelemaan samankaltaisia kokonaisuuksia ja tunnistamaan tekstistä haettuja arvoja. Materiaalitieteissä NER on vielä alkutekijöissään puutteellisten tietokantojen takia. Laajana tieteenalana materiaalitieteiden erilaisia entiteettejä on hyvin paljon, ja niiden kartoittaminen on haastavaa, koska olemassaolevia entiteettejä on vähän käytettäväksi vertausarvoina.[16]

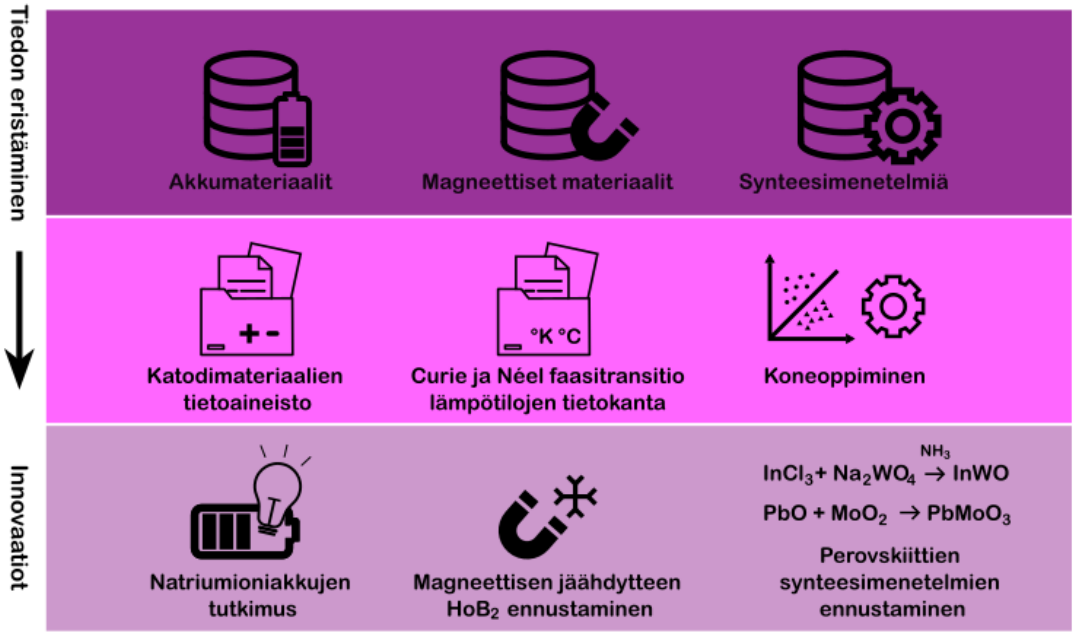
Koneoppimiseen perustuva NER pystyy tunnistamaan entiteettejä erilaisiin ominaisuuksiin perustuvan datan avulla. Kielimallit kykenevät niille opetettujen ennalta merkittyjen entiteettien avulla vertailemaan sanojen ja lauseiden ominaisuuksia ja määrittämään niille oikeat tunnisteet. Varsinkin kaksisuuntaiset kielimallit, kuten BERT pystyvät vertaamaan käsiteltävää sanaa kaikkiin muihin sanoihin ja niiden ominaisuuksiin ja näin määrittämään tunnisteen. Näiden kielimallien koulutus vaatii kuitenkin korkealaatuista asiantuntijoiden luomaa dataa, jossa on huomioitu yksityiskohdat tunnisteiden luokittelussa.[16]

Haasteita nimettyjen entiteettien tunnistamisessa on myös tiedon kontekstiin ja merkintätapoihin liittyen. Materiaaleilla voi olla vaihtelevia nimeämistapoja sekä jotkin materiaalit voivat koostua useista eri sanoista. On siis haastavaa kehittää materiaalitieteiden eri erikoisalojen välillä yleisesti toimivia tiedon eristämisen työkaluja. Työkalujen kehittäminen vaatii usein ohjatun oppimisen lähestymistapoja.[19]

Entiteettien tunnistamisen jälkeen niiden välisiä suhteita ja yhteyksiä voidaan tutkia. Entiteettien välisiä suhteita voi olla lukuisia, esimerkiksi entiteettien ominaisuudet ja kuinka usein ne esiintyvät toisten entiteettien yhteydessä. Tuntemattomia entiteettejä verrataan opetettuihin kaavoihin ja voidaan näin luoda uusia suhteita eri entiteettien välillä.[16] Näiden vaiheiden jälkeen entiteettien tarjoamaa dataa voidaan hyödyntää tiedonrikastuksessa tai tiedonlouhinnassa, eli aiemmin kuvatuilla koneoppimismenetelmillä, eri tarkoituksiin, kuten uusien materiaalien havaitsemiseen.

Tiedon eristämisen prosessissa on mainitsemisen arvoista huomioida tiettyjä seikkoja: onko tieto eristetty virheettömästi, onko tieto kirjattu ylös oikealla tavalla ja onko saatu eristettyä tarpeeksi tarkkaa ja yksityiskohtaista tietoa. Jos eristetty tieto ei ole tarpeeksi laadukasta, ei sen eristäminen ole kannattavaa. Datan paikkansapitävyys heikkenee tiedon eristämisen vaiheiden aikana, joten on tärkeää, että prosessi on mahdollisimman sujuva. Datan määrään ja laatuun vaikuttavat mm. käytetyt tiedon eristämisen menetelmät, artikkeleissa käytetyt epätäydelliset viittaukset aiempaan dataan ja epätyypillisten termien käyttö. Pienempien tietoaineistojen käyttö voi tarjota tarkempia tuloksia, mutta suurempia aineistoja käsittelemällä saadaan enemmän dataa.[16]

Kieliteknologioiden hyödyntämisestä materiaalitieteissä on jo jonkin verran esimerkkejä tietokantojen sekä tietoaineistojen kokoamisesta koneoppimista hyödyntäviin oivalluksiin. Seuraavia esimerkkejä havainnollistetaan kuvassa 6. Gou ym. laativat työnkulun katodimateriaalien eristämiseen ja loivat tietoaineiston, jota voidaan hyödyntää natriumioniakkujen tutkimuksessa ja kehityksessä.[19] Court ja Cole loivat tietokannan magneettisten materiaalien Curien ja Néelin faasitransitiolämpötiloista, jota hyödyntämällä Castro ym. löysivät magneettisen jäähdytteen.[16,20,21] Kim ym. tutkivat epäorgaanisten materiaalien synteesimenetelmiä neuroverkkoja hyödyntämällä ja onnistuivat määrittämään kahdelle perovskiitille prekursorit eli lähtöaineet hyödyntämällä koulutusdataa, joka oli julkaistu ennen niiden ensimmäistä todistettua syntetisointia.[16,22]



Kuva 6. Esimerkkejä tiedon eristämisen hyödyntämisestä materiaalitieteen sovelluksissa.

4 Eri tavoin koulutetut materiaalitekniikan kielimallit

Kielimallien menestys yleisen tason tehtävissä on johtanut alakohtaisten ja hienosäädettyjen kielimallien kehitykseen. Suuria kielimalleja voidaan jatkokouluttaa ja hienosäätää erikoistumaan tietyn tieteenalan tekstin piirteisiin, konteksteihin ja NLP-tehtäviin. Tieteellisten tekstien käsittelyyn on koulutettu kielimalleja, kuten SciBERT[23], joka ymmärtää paremmin tieteellisen tekstin kielioppia, termistöä ja rakenteita sen koulutuksessa käytetyn tieteellisen korpuksen ansiosta. Ne ovat myös materiaalitieteissä osoittautuneet tehokkaammiksi erilaisissa luonnollisen kielen käsittelyn tehtävissä.[24]

Alan monimuotoisen luonteen vuoksi on kehitetty erilaisia hienosäädettyjä materiaalikielimalleja, jotka suoriutuvat paremmin alan tyypillisistä entiteetteihin liittyvistä tehtävistä. Näitä ovat esimerkiksi MatSciBERT[25], MatBERT[26], MaterialsBERT[27], MaterialBERT[28], BatteryBERT[29] ja OpticalBERT[30].

GPT-malleilla on onnistuttu saavuttamaan BERT-malleja vastaavia tuloksia vähemmän dataa vaativilla koulutusaineistoilla. Näissä tuloksissa on kuitenkin huomioitava GPT-mallien generatiivinen luonne, joka aiheuttaa mallin keksivän tuloksia tyhjästä ja antavan vääriä vastauksia oikeina.[17] GPT-mallit eivät ole todennäköisesti ideaalisia tiedon eristämisen tehtäviin niiden generatiivisen, suljetun ja kaupallisen luonteen vuoksi, mutta niillä on tieteidenvälisessä alassa potentiaalia mm. avustaa matalan tason tiedon tiivistämisessä.[18,31]

Kielimallien tehokkuutta ja luotettavuutta voidaan arvioida tutkimalla sisäistä tarkkuutta (engl. precision), herkkyyttä (engl. recall) ja niiden välistä tasapainotettua keskiarvoa F1-arvoa (engl. F1-score).[19]

$$Sisäinen\ tarkkuus = \frac{OP}{OP + VP} \tag{1}$$

$$Herkkyys = \frac{OP}{OP + VN} \tag{2}$$

$$F1 = \frac{2 \times Sisäinen\ tarkkuus \times Herkkyys}{Sisäinen\ tarkkuus + Herkkyys} \tag{3}$$

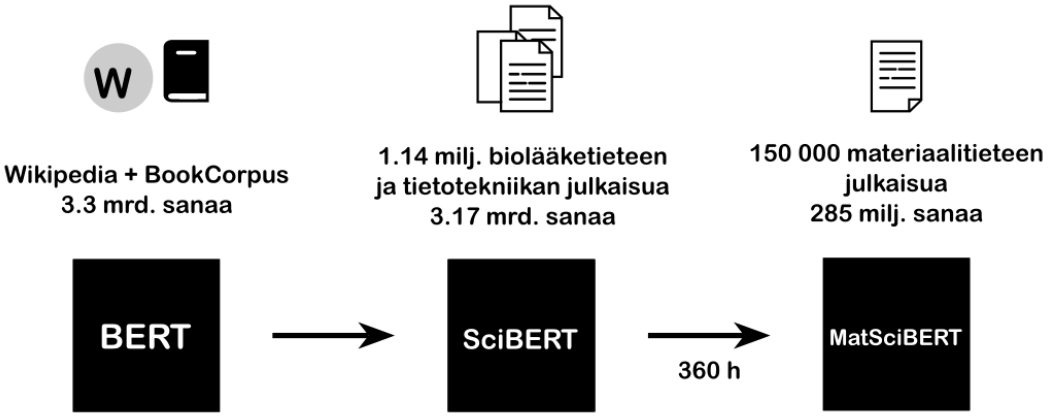
jossa OP on oikeat positiiviset, VP väärät positiiviset ja VN väärät negatiiviset. Sisäinen tarkkuus kuvaa positiivisiksi luokiteltujen entiteettien todellista määrää ja herkkyys mallin kykyä luokitella oikein kaikki oikeasti positiiviset entiteetit.

4.1 MatSciBERT

MatSciBERT on Delhin teknillisessä korkeakoulussa kehitetty materiaalitekniikan kielimalli. Se esiteltiin ensimmäisen kerran syyskuussa 2021 ensimmäisenä materiaaleihin erikoistuneena kielimallina sekä kokonaisuudessaan toukokuussa 2022 julkaistussa artikkelissa.[25] MatSciBERT jatkokoulutettiin SciBERTistä käyttäen Elsevier-kustantamon julkaisuista koottua materiaalitieteen tietokantaa. Kuvassa 7 havainnollistetaan kielimallin siirto-oppimista ja käytettyjen tietokantojen kokoa.

Koulutuksessa käytetty 285 miljoonan sanan materiaalitiedekorpus koostui sekä kokeellista että laskennallista dataa sisältävistä julkaisuista. Keraamit ja epäorgaaniset lasit kattoivat 40 % korpuksen sanoista. Metalliseokset, metalliset lasit ja sementti kattoivat jokainen 20 %:n osuuden sanoista. Nämä materiaaliluokat käsittivät myös lämpösähkötekniikan materiaaleja, nanomateriaaleja, polymeerejä ja biomateriaaleja. Korpuksen tokenoinnissa hyödynnettiin SciBERTin sanastoa ja painokertoimia. Materiaalitieteisiin erikoistunut tokenointi voisi parantaa kielimallin toimintaa. Olemassa olevan sanaston käyttöä puoltaa kuitenkin kielimallien kyky ymmärtää kontekseja vajavaisesta tokenoinnista huolimatta.[25]

MatSciBERTin toimintaa kokeiltiin kolmessa eri tiedon eristämisen tehtävässä, joita olivat NER, entiteettien yhteyksien luokittelu ja abstraktien luokittelu. Apuna käytettiin tietoaineistoja, jotka asiantuntijat ovat annotoineet, eli merkinneet ja luokitelleet oikein. Mallin kykyä saada vastaavat tulokset verrattiin myös SciBERTiin, BERTiin ja tietoaineistojen laatijoiden neuroverkkomalleilla saavuttamiin parhaisiin tuloksiin. MatSciBERT saavutti kaikista tehtävistä paremman F1-arvon kuin aiemmin mainitut mallit.[25]



Kuva 7. MatSciBERTin siirto-oppiminen ja tietokantojen suuruus.[14,23,25]

Nimettyjen entiteettien tunnistuksessa kaikista kielimalleista käytettiin kolmea rakenteellisesti hieman eroavaa versiota (Linear, CRF, BiLSTM-CRF). Tietoaineistoina toimivat 3 eri aineistoa. SOFC (Solid Oxide Fuel Cells) ja SOFC-Slot eli tietoaineistot kiinteistäoksidipolttokennoista, jotka koostuvat samasta datasta, mutta aiemmassa on 4 annotoitua entiteettityyppiä ja jälkimmäisessä 16. Matscholar-tietoaineistossa on 7 annotoitua entiteettityyppiä, esimerkiksi epäorgaaninen materiaali, materiaalin ominaisuus ja materiaalin sovellus.[25] Taulukkoon 1 on valikoitu Gupta ym. artikkelin tuloksien jokaisesta tehtävästä parhaiten suoriutuneen MatSciBERT-version F1-arvo verrattuna vastaavaan SciBERTin ja BERTin tulokseen.

Entiteettien yhteyksien luokittelussa käytettiin materiaalien synteesimenetelmä tietoaineistoa, jossa synteesimenetelmille on luokiteltu vaiheet ja niihin liittyvät entiteetit. Abstraktien luokittelussa tietoaineistona toimi 1500 abstraktin tietoaineisto, joka on luokiteltu sen mukaan, liittyykö abstrakti lasimateriaaleihin vai ei.[25] Taulukossa 1 on BERT-mallien yhteyksien luokittelun F1-arvo ja abstraktien luokittelun tarkkuus eli kuinka monta abstraktia malli luokitteli oikein prosentteina.

MatSciBERT ja kaikki koulutuksessa sekä testauksessa käytetty koodi ovat avoimesti ladattavana verkossa omaa käyttöä tai jatkokouluttamista varten.[25]

Taulukko 1. Eri tiedon eristämisen tehtävissä, eri tietoaineistoilla ja eri kielimalleilla saavutettuja tuloksia.[25]

Tehtävä/tietoaineisto	Evaluointiarvo	MatSciBERT	SciBERT	BERT
NER - SOFC-Slot	F1	**65.95 ± 2.53**	61.68 ± 1.42	55.44 ± 1.97
NER - SOFC	F1	**82.39 ± 1.23**	81.07 ± 0.93	78.93 ± 1.62
NER - MatScholar	F1	**86.38 ± 0.49**	85.04 ± 0.77	84.07 ± 0.19
Entiteettien yhteyksien luokittelu - synteesimenetelmät	F1	**89.02 ± 0.27**	87.22 ± 0.58	85.40 ± 1.45
Abstraktien luokittelu - Lasimateriaalit	Tarkkuus	**96.22 ± 0.16**	93.44 ± 0.57	93.89 ± 0.68

4.2 MatBERT

MatBERT on Berkeley-yliopistossa kehitetty materiaalitekniikan kielimalli, joka esiteltiin huhtikuussa 2022 julkaistussa artikkelissa.[26] MatBERT koulutettiin materiaalitieteen tieteellisillä julkaisuilla, joita valittiin aluksi satunnaisesti 2 miljoonaa. Tokenoinnissa käytettiin kahta WordPiece-tokenoijaa, joista toinen huomioi kirjainten ja merkkien kokoa. WordPiece-tokenoija paloittelee sanat pieniin osiin, mikä parantaa kielimallin tuntemattomien ja vaikeiden sanojen käsittelyä. Koulutuksessa, joka kesti noin kuukauden, käytettiin kaikkiaan 8.8 miljardin tokenin materiaalikorpusta.

Trewartha ym. testasivat MatBERTiä kolmella eri tietoaineistolla nimettyjen entiteettien tunnistamisessa. Testaamiseen käytettiin versiota, joka ei huomioi merkkien kokoa. MatBERTin tuloksia verrattiin SciBERTin, BERTin ja BiLSTM-mallin (Bidirectional Long Short Term Memory Network) tuloksiin. Yhtenä tietoaineistoista toimi aiemmin esitelty MatScholar-tietoaineisto. Puolijohteiden seostamisessa käytettyjen materiaalien tietokanta koostuu 455 abstraktista, joista on annotoitu 3 eri entiteettiä. Kultananopartikkelien synteesiin liittyvistä 73 artikkelista koostetussa tietokannassa entiteetit luokiteltiin kahteen tyyppiin.[26]

Jokaisesta testatusta tietoaineistosta arvioitiin jokaisen entiteetin tyyppikohtaiset F1-arvot. MatBERT saavutti yhtä entiteettityyppiä lukuun ottamatta parhaan tuloksen. SciBERT oli käytetyistä kielimalleista toiseksi paras. Taulukossa 2 on Trewartha ym. tuloksien pohjalta laskettu keskiarvot jokaisen tietoaineiston F1-arvoista.

MatBERT ja sen koulutuksessa käytetty koodi ovat avoimesti ladattavana verkossa.[26] Sitä on käytetty myös muiden materiaaliluokkien eristämiseen. Zhang ym. kehittivät mallia eteenpäin vuonna 2024 julkaisemassaan artikkelissa[32], ja käyttivät sitä nimettyjen entiteettien eristämiseen kokoamastaan perovskiitti-materiaalien tietoaineistosta. Muokattu MatBERT saavutti F1-arvon 90.8, MatBERT 89.7 ja SciBERT 87.3.

Taulukko 2. MatBERTillä eri tietoaineistoilla nimettyjen entiteettien tunnistuksessa saatujen F1-arvojen keskiarvot.[26]

	MatBERT	SciBERT	BERT	BiLSTM
MatScholar	**87.00**	85.57	82.29	82.71
Seostetut puolijohteet	**72.00**	70.00	64.33	63.33
Kultananopartikkelit	**79.50**	59.50	41.00	71.50

4.3 MaterialsBERT

MaterialsBERT on Yhdysvalloissa Georgian teknillisessä yliopistossa kehitetty materiaalitekniikan kielimalli, jonka Shetty ym. esittelivät huhtikuussa 2023 julkaistussa artikkelissa.[27] MaterialsBERT koulutettiin käyttäen pohjana biolääketieteen kielimallia PubMedBERT, joka on koulutettu 14 miljoonalla abstraktilla ja kokonaisilla julkaisuilla PubMed-tietokannasta. Se valittiin perustaksi, koska biolääketiede kattaa paljon samoja entiteettejä kuin materiaalitieteet. Hienosäätämiseen käytettiin 2.4 miljoonaa materiaalitieteen abstraktia ja siinä kesti 90 tuntia. Tokenointiin käytettiin WordPiece-tokenointia.

MaterialsBERTin 2.4 miljoonan korpuksesta n. 650 000 abstraktia liittyivät polymeereihin. Shetty ym. laativat myös polymeeritietoaineiston annotoimalla 750 abstraktia polymeereistä. Tietoaineiston avulla koulutettiin NER-malli, jolla eristettiin n. 300 000 polymeerien ominaisuuksiin liittyvää entiteettiä. Tämä vei 60 tuntia aikaa verrattuna PoLyInfo-tietokantaan, joka on n. 500 000 entiteettiä kattava asiantuntijoiden manuaalisesti monien vuosien aikana kokoama polymeeritietokanta. Kielimallilla eristetty tietokanta vaatii myös asiantuntijoiden annotointia, mutta työnmäärä on merkittävästi pienempi.[27]

Kielimallia arvioitiin nimettyjen entiteettien tunnistamisessa viidessä tietoaineistossa, joita olivat työssä koottu polymeerikorpus, aiemmissa osioissa mainitut MatScholar-tietoaineisto ja materiaalien synteesimenetelmä tietoaineisto sekä kaksi suurta kemiallisten entiteettien aineistoa. MaterialsBERTiä verrattiin muihin BERT-kielimalleihin, ja se sai parhaan F1-arvon kolmesta tietoaineistosta.[27] Taulukossa 3 on esitetty Shetty ym. artikkelissa saadut MaterialsBERTin, PubMedBERTin ja MatBERTin F1-arvot eri tietoaineistoista.

MaterialsBERT, sen koulutuksessa käytetty koodi sekä polymeeritietoaineisto ovat avoimesti ladattavissa netistä. Työssä eristettyä polymeerien ominaisuuksien tietokantaa varten tehtiin nettisivu yksinkertaisella graafisella käyttöliittymällä tietokannan tutkimista varten.[27]

Taulukko 3. Kolmen kielimallin NER F1-arvot viidestä eri tietoaineistosta.[27]

	MaterialsBERT	PubMedBERT	MatBERT
Polymeerit	**66.4**	65.8	65.2
MatScholar	86.0	85.0	**86.2**
Synteesimenetelmät	**68.6**	67.6	68.2
ChemDNER	69.2	**70.2**	69.2
ChemRxn	**71.4**	63.6	62.4

5 Yhteenveto ja pohdinta

Tiedon eristäminen materiaalitieteissä on haastavaa johtuen alan monimuotoisesta luonteesta. Hyödyllisen tiedon eristäminen vaatii tarkkoja kielimalleja, joiden koulutus on kuitenkin vaikeaa laadukkaasti annotoidun datan vähäisyyden takia. Materiaalitieteiden alalla yleispätevästi toimivan kielimallin kehitys on haastavaa, koska yksittäiset materiaalitieteiden erikoisalat vaativat usein erilaista osaamista ja toimintakykyä malleilta. Käsiteltävät entiteetit, menetelmät, merkitsemistavat ja muut käytänteet vaihtelevat paljon.

Materiaalitekniikan kielimalleja onkin kehitetty lyhyessä ajassa useita. Näillä kielimalleilla on usein hieman eroavia käyttötarkoituksia, mutta niiden testaamisessa on käytetty monesti samoja tietoaineistoja, joilla on saavutettu samankaltaisia tuloksia. Esimerkiksi MatSciBERT, MatBERT ja MaterialsBERT saivat kaikki MatScholar-tietoaineistosta F1-arvon välillä 86–87 %. Käytetyt arviointimenetelmät vaihtelevat tietysti hieman, mutta tulokset ovat hyvin samankaltaisia ottaen huomioon eri mallien kehityksen vaatimat resurssit. Tiedon eristämisen ytimessä on aiemman tiedon päälle rakentaminen, mikä olisi hyvä huomioida myös materiaalitekniikan kielimalleja kehitettäessä. Tätä hyödynnetään kuitenkin jo BERTin sekä SciBERTin ja PubMedBERTin jatkokouluttamisessa materiaalitieteen tarkoituksiin.

Vaikka materiaalitekniikan kielimalleilla on saatu parempia F1-arvoja, on huomion arvoista pohtia perustelevatko muutaman prosentin korkeammat tulokset erikoistuneiden mallien työlästä kehitystä. Suuria korpuksia käsitellessä hiemankin korkeampi tulos voi toki merkitä tuhansia kallisarvoisia eristettyjä entiteettejä. Täytyy kuitenkin huomioida, että F1-arvo on tasapainotettu keskiarvo eikä siitä voida kertoa suoraan eristettyjen entiteettien määrää. Myös testauksessa käytetyt tietoaineistot ovat hyvin pieniä, joten on haastavaa arvioida kielimallien todellista kykyä ennen laajempaa tutkimusta suuremmilla aineistoilla.

Materiaalitieteiden julkaisujen dataa on usein vaikea eristää. Materiaalitieteiden erikoisalojen välillä tulisi pyrkiä yhtenäistämään tiedon merkintä- ja jakamistapoja helposti käsiteltäviksi tiedon eristys tarkoituksiin. Tiedon eristämisen työkalut kuten suuret kielimallit vaativat kattavaa tietoteknistä osaamista, joka ei ole tyypillistä useimmille materiaalitieteilijöille. Kielimallien kehityksen ja saavutettavuuden kannalta olisi hyödyllistä pyrkiä tekemään niiden käyttäminen yksinkertaisemmaksi. Tietoaineistojen, tietokantojen ja kielimallien kehittyessä yhden alan tutkimustuloksista eristettyä tietoa voitaisiin hyödyntää helposti muiden alojen tukena. Näin materiaalitieteiden monimuotoisuuden heikkoudesta voisi kehittyä vahvuus.

Lähteet

1. Deng, L. & Liu, Y. *Deep learning in natural language processing*. Springer, 1-22 (2018).

2. Eisenstein, J. *Introduction to natural language processing*. MIT press, 1-10 (2019).

3. Hu, F. & Hao, Q. *Intelligent sensor networks: the integration of sensor networks, signal processing and machine learning*. Taylor & Francis, 3-29 (2012).

4. Murphy, K. P. *Probabilistic machine learning: an introduction*. MIT press, 1-29 (2022).

5. Sabharwal, N. & Agrawal A. *Hands-on Question Answering Systems with BERT: Applications in Neural Networks and Natural Language Processing*. Springer, 1-14 (2021).

6. Erkan, A. & Güngör, T. Analysis of Deep Learning Model Combinations and Tokenization Approaches in Sentiment Classification. *IEEE Access* 11, 134951–134968 (2023).

7. Martinez, A. R. Part-of-speech tagging. *WIREs Computational Stats* 4, 107–113 (2012).

8. Li, J. ym. A survey on deep learning for named entity recognition. *IEEE transactions on knowledge and data engineering* 34.1, 50-70 (2020).

9. Callister, W. D. & Rethwisch, D. G. *Materials Science and Engineering: An Introduction: SI Version*. Wiley, 393 (2020).

10. Dongare, A. D. ym. Introduction to artificial neural network. *International Journal of Engineering and Innovative Technology (IJEIT)* 2.1, 189-194 (2012).

11. Vaswani, A. ym. “Attention is all you need.” *Advances in neural information processing systems* 30 (2017).

12. Naveed, H. ym. A Comprehensive Overview of Large Language Models. Ennakkopainos arXiv:2307.06435 (2024).

13. Brown, T. ym. Language models are few-shot learners. *Advances in neural information processing systems* 33, 1877-1901 (2020).

14. Devlin, J. ym. Bert: Pre-training of deep bidirectional transformers for language understanding. Ennakkopainos arXiv:1810.04805 (2018).
15. Zhong, X. ym. Explainable machine learning in materials science. *npj Comput Mater* 8, 204 (2022).
16. Olivetti, E. A. ym. Data-driven materials research enabled by natural language processing and information extraction. *Applied Physics Reviews* **7**, 041317 (2020).
17. Choi, J. & Lee, B. Accelerating materials language processing with large language models. *Commun Mater* 5, 13 (2024).
18. Lei, G. ym. Materials science in the era of large language models: a perspective. *Digital Discovery* (2024).
19. Gou, Y. ym. A document-level information extraction pipeline for layered cathode materials for sodium-ion batteries. *Sci Data* 11, 372 (2024).
20. Court, C. J. & Cole, J. M. Auto-generated materials database of Curie and Néel temperatures via semi-supervised relationship extraction. *Sci Data* 5, 180111 (2018).
21. Castro, P. B. D. ym. Machine-learning-guided discovery of the gigantic magnetocaloric effect in HoB2 near the hydrogen liquefaction temperature. *NPG Asia Mater* 12, 35 (2020).
22. Kim, E. ym. Inorganic Materials Synthesis Planning with Literature-Trained Neural Networks. *J. Chem. Inf. Model.* 60, 1194–1201 (2020).
23. Beltagy, I. ym SciBERT: A pretrained language model for scientific text. Ennakkopainos arXiv:1903.10676 (2019).
24. Song, Y. ym. MatSci-NLP: Evaluating Scientific Language Models on Materials Science Language Tasks Using Text-to-Schema Modeling. Ennakkopainos arXiv:2305.08264 (2023).

25. Gupta, T. ym. MatSciBERT: A materials domain language model for text mining and information extraction. *npj Comput Mater* 8, 102 (2022).
26. Trewartha, A. ym. Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science. *Patterns* 3, 100488 (2022).
27. Shetty, P. ym. A general-purpose material property data extraction pipeline from large polymer corpora using natural language processing. *npj Comput Mater* 9, 52 (2023).
28. Yoshitake, M. ym. MaterialBERT for natural language processing of materials science texts. *Science and Technology of Advanced Materials: Methods* 2, 372–380 (2022).
29. Huang, S. & Cole, J. M. BatteryBERT: A Pretrained Language Model for Battery Database Enhancement. *J. Chem. Inf. Model.* 62, 6365–6377 (2022).
30. Zhao, J. ym. OpticalBERT and OpticalTable-SQA: Text- and Table-Based Language Models for the Optical-Materials Domain. *J. Chem. Inf. Model.* 63, 1961–1981 (2023).
31. Hira, K. ym. Reconstructing Materials Tetrahedron: Challenges in Materials Information Extraction. *Digital Discovery* 3, 1021–1037 (2024).
32. Zhang, J. ym. Named entity recognition in the perovskite field based on convolutional neural networks and MatBERT. *Computational Materials Science* 240, 113014 (2024).