Tiivistelmä

Tallennettavan hoitotiedon määrä sairaaloissa lisääntyy jatkuvasti sähköisten potilastietojärjestelmien yleistymisen myötä. Suuri osa hoitoalan ammattilaisten kirjaamasta hoitodokumentaatiosta tallennetaan vapaamuotoisena tekstinä erilaisina kliinisinä muistiinpanoina ja hoitokertomuksina. Jokainen hoitojakso dokumentoidaan säännöllisesti kirjattavilla hoitokertomuksilla sekä hoitojakson päättyessä laadittavalla tiivistelmämuotoisella hoitopalautteella. Vaikka sähköiset potilasjärjestelmät ovat oivallisia tällaisen tiedon hallinnoinnissa, ei tallennetun tiedon hyödyntämisen täyttä potentiaalia paremman hoidon toteuttamisessa tai tutkimus- ja opetuskäytössä ole vielä saavutettu. Hoitoalalle suunnatun kieliteknologian kehittyminen on kuitenkin mahdollistanut tekstimuotoisten potilastietojen tehokkaamman käyttämisen, muun muassa tiedonhaussa, terminologian kehittämisessä, tiedon uuttamisessa, kielen kääntämisessä ja tiivistelmien tuottamisessa.

Tässä tutkielmassa tarkastellaan useita automatisoituja menetelmiä, joilla arvioidaan tekstin semanttisen samankaltaisuutta sekä uutetaan tietoa sähköisten potilastietojärjestelmien vapaamuotoisesta tekstidatasta. Tutkimuksen pääpaino on automatisoitujen tiivistelmien tuottamisessa hoitojaksojen aikana kirjoitetuista hoitokertomuksista. Tutkimuksen teemana ovat pieniresurssiset menetelmät, joita hyödyntämällä voidaan vähentää manuaalisesti luotujen tietämyskantojen ja koulutusdatan tarvetta. Tämän vuoksi tekstien semanttista samankaltaisuutta arvioidaan suurista kliinisistä tekstiaineistoista johdetuilla sanojen tilastollisilla jakaumilla, jotka esitetään vektorimuotoisina malleina (engl. distributional semantic models). Näitä malleja hyödynnetään myös hoitokertomusten tiivistelmien tuottamisessa.

Tutkielmassa esitettyjä menetelmiä verrataan hoitoalan asiantuntijoiden tuottamiin analyyseihin. Nämä tulokset osoittavat, että jakaumalliset mallit ovat lupaava pieniresurssinen menetelmä automatisoituun semanttisen samankaltaisuuden määrittämiseen kliinisistä tekstikorpuksista. Tutkielmassa esitetyt parannukset perinteisiin menetelmiin nähden pohjautuvat vektoriesitysten kykyyn kuvata semanttisia piirteitä aiempaa kattavammin. Nämä vektoriesitykset perustuvat joko useisiin eri tekstiaineistoilla ja piirteillä koulutettuihin semanttisiin malleihin, tai malleihin, jotka käyttävät erillisiä vektoriesityksiä yhden sanan eri merkityksille. Lisäksi tutkimuksessa hyödynnetään hoitojaksojen raportoinnissa käytettävää metadataa semanttisten mallien kouluttamiseen, minkä tavoitteena on parantaa mallien soveltuvuutta tiedonhakuun kliinisistä tekstiaineistoista. Tutkimuksen tulokset osoittavat, että tällaiset menetelmät soveltuvat hyvin kliiniseen tiedonhakuun.

Tutkimuksessa myös osoitetaan että Random Indexing -menetelmää voidaan laajentaa siten, että sanojen eri merkityksille tuotetaan omat vektoriesityksensä. Neuroverkkoihin perustuva Word2vec-menetelmä suoriutuu kuitenkin perinteistä Ran-

dom Indexing -menetelmää paremmin useissa samankaltaisuuden arviointitehtävissä, kun valitut parametrit ovat vertailukelpoisia ja koulutus on toteutettu samalla aineistolla. Lisäksi tutkielmassa on tarkasteltu useita kliinisen tekstin tilastollisia ominaisuuksia ja niiden kykyä ilmaista virkkeiden oleellisuutta tiivistelmien tuottamisessa sekä semanttisten mallien hyödyntämistä erillisten hoitojaksojen ja niiden hoitopalautteiden samankaltaisuuden määrittämisessä. Tuotettujen tiivistelmien laatua on arvioitu eri näkökulmista yhdessä hoitoalan asiantuntijoiden kanssa sekä automaattisin menetelmin vertailemalla tuotettuja tiivistelmiä hoitojaksojen alkuperäisiin hoitopalautteisiin. Asiantuntija-arvioiden sekä automaattisten arvioiden vertailu osoittaa, että näiden välillä on vahva korrelaatio ja että hoitopalautteita voidaan hyödyntää mallitiivistelminä.