Tietojenkäsittelytiede **28** Joulukuu 2008 sivut 4–14
Toimittaja: Jorma Tarhio

Kieliteknologiasta ratkaisu potilasdokumentaation hyödynnettävyyteen

Hanna Suominen
Turku Centre for Computer Science (TUCS)
Turun yliopisto, Informaatioteknologian laitos
hajasu@utu.fi

Tiivistelmä

Teho-osaston hoitotyössä kirjataan paljon vapaamuotoista tekstiä ja näiden dokumenttien rakenne on epäyhtenäinen. Tämä hankaloittaa kirjattavan tiedon tehokasta jatkokäyttöä ja työntekijöiden välistä tiedonkulkua. Kieliteknologia voi tarjota uudenlaisia mahdollisuuksia helpottaa vapaamuotoisen potilasdokumentaation tuottamista ja sen hyödyntämistä potilaiden hoitoon liittyvässä päätöksenteossa teho-osastoilla. Erityisesti tiedonhakuun sovelletut luokittelu-, relevanssinarviointi- ja aihesegmentointialgoritmit ovat tuottaneet jatkotutkimusta ja pilotointia rohkaisevia tuloksia.

1 Johdanto

Potilasasiakirjojen sähköistyminen on tarjonnut uusia mahdollisuuksia tietojen kirjaamiseen potilasasiakirjoihin ja tallennetun tiedon hyödyntämiseen. Kuitenkin potilasasiakirjoihin vapaamuotoisena tekstinä kirjattavan tiedon tehokas jatkokäyttö on nykyisellään vaikeaa, sillä esimerkiksi tietyn asian kannalta olennaisten katkelmien poimiminen on vapaamuotoisesta tekstistä työlästä. Koska vapaamuotoisen tekstin käyttö on numeeriseen tai rakenteiseen tietoon verrattuna erilaisen ilmaisuvoimansa vuoksi potilasdokumentaatiossa välttämättömyys [12, 29], on sen hyödyntämiseen alettu kehittää kieliteknologiaa. Tämä työ on paitsi terveydenhuollon ammattilaisten ja terveysteknologiatoimijoiden myös akateemisten tutkijoiden kasvavan kiinnostuksen kohde: kieliteknologia nähdään yhtenä vastauksena tarpeeseen tukea vapaamuotoisen tekstin tuottamista ja sen hyödynnettävyyttä potilaan hoitoon liittyvässä päätöksenteossa, terveyspalvelujen tuottajien hallinnollisissa tehtävissä sekä lääketieteellisessä tutkimuksessa ja koulutuksessa.

Potilasasiakirjoihin soveltuvaa kieliteknologiaa on kehitetty viime aikoina esimerkiksi tunnistamaan potilasasiakirjojen vapaatekstiosioiden perusteella tietyn sairauden omaavat potilaat [7, 17, 18] ja täyttämään erilaisia potilastietolomakkeita poimimalla automaattisesti sopivat tekstikatkelmat potilasasiakirjoista [3, 14, 22, 32]. Lisäksi tätä teknologiaa on hyödynnetty muun muassa keräämällä potilasasiakirjoista hoitotoimenpiteisiin liittyviä haittavaikutuksia [19] ja sairauksien hoitoon liittyvää tietoa [31]. Vapaamuo-

OHJE KIRJAPAINOLLE: B5-arkin vasen- ja yläreuna kohdistetaan A4-arkin vasempaan ja yläreunaan. Näin pitäisi marginaaliksi tulla taitteen puolella noin 33 mm ja muualla noin 22 mm.

toiseen potilasdokumentaatioon perustuvia terminologiamalleja on muodostettu patologian alalle [33]. Eräät kieliteknologiset sovellukset ovat jo rutiinikäytössä sairaaloissa. Esimerkkejä tällaisista sovelluksista ovat New York Presbyterian Hospitalissa kliinisten raporttien jäsentämiseen ja koodaamiseen käytettävä MedLEE [13] sekä Rochesterin Mayo Clinicalla kliiniset diagnoosit automaattisesti määrittävä Autocoder [16]. Edellä kuvatuista menestyksekkäistä kieliteknologian sovellusesimerkeistä huolimatta on muistettava, että harvinaisilla kielillä, kuten suomeksi, kirjatun potilasdokumentaation analysointiin tarkoitettujen ohjelmien kehitys on vasta aluillaan.

Väitöstutkimukseni tavoitteena on tutkia, miten kieliteknologian menetelmiä voidaan soveltaa suomenkielisen vapaamuotoisen potilasdokumentaation sisällön jatkokäsittelyyn ja kehittää yhteistyössä muiden tutkijoiden kanssa vapaamuotoisen potilasdokumentaation sisällön hyödyntämistä tukevia kieliteknologisia komponentteja. Väitöstutkimus edistää sovellusalueensa lisäksi myös yleisemmin kieliteknologian tutkimusta ja suomenkieliselle tekstille soveltuvien ohjelmien kehitystä. Onnistuessaan kehitettävät kieliteknologiset komponentit voivat helpottaa hoitohenkilökunnan työtä ja tukea heidän päätöksentekoaan. Tutkimusaineistoksi on valittu teho-osaston hoitotyön potilasdokumentaatio, koska teho-osastoilla työskentelevien hoitajien päätöksenteko on eri maissa varsin samankaltaista [10]. Näin ollen voidaan olettaa myös teho-osastoilla hoitajien päätöksenteon tueksi tarvittavat sovellukset kansainvälisesti samankaltaisiksi. Teen tutkimustani Turun tietotekniikan tutkimus- ja koulutuskeskuksessa (Turku Centre for Computer Science, TUCS) ja Turun yliopiston informaatioteknologian laitoksella osana Tekesin FinnWell-ohjelmaan kuuluvaa Potilasasiakirjojen tekstin louhinta (Louhi) -hanketta [11].

Esitän seuraavaksi tutkimustuloksiani vuosilta 2005–2008. Aloitan kuvaamalla luvussa 2 hoitotyön vapaamuotoista potilasdokumentaatiota teho-osastolla. Luvussa 3 pohdin kieliteknologialla saavutettuja etuja sekä kieliteknologian kehittämiseen ja käyttöön liittyviä riskejä. Sitten etenen luvussa 4 potilasasiakirjojen vapaatekstiosioiden hyödynnettävyyden tukemiseen jakamalla teksti automaattisesti aiheen mukaisiksi kappaleiksi ja nimeämällä näin muodostuneiden tekstikatkelmien aiheet. Lopuksi käsittelen jatkotutkimusta luvussa 5. Vuosia 2005 ja 2006 koskeva osa perustuu artikkeliini [23].

2 Hoitotyön vapaamuotoinen potilasdokumentaatio teho-osastolla

Tehohoito määritellään vaikeasti sairaiden potilaiden hoidoksi, jossa potilasta tarkkaillaan keskeytymättä ja hänen elintoimintojaan valvotaan ja tarvittaessa pidetään yllä erikoislaittein hoidon tavoitteen ollessa voittaa aikaa perussairauden hoitamiseen torjumalla hengenvaara [1]. Potilaiden jatkuvan tarkkailun vuoksi tehohoidossa kertyy paljon potilasdokumentaatiota.

Elektronisiin potilastietojärjestelmiin siirtyminen on käynnissä parhaillaan Suomen teho-osastoilla. Esimerkiksi Turun yliopistollisen keskussairaalan aikuisten teho-osastolla elektroninen potilastietojärjestelmä on otettu käyttöön keväällä 2002. Käytettäessä elektronista potilastietojärjestelmää hoitohenkilökunta täydentää muun muassa hengityskoneesta ja valvontalaitteista automaattisesti kertyvää numeerista tai strukturoitua tietoa. Mer-

kittävä osa tästä täydentävästä tiedosta on vapaamuotoista tekstiä.

Tuloraportit, hoidon aikana kirjatut muistiinpanot ja uloskirjausraportit muodostavat hoitotyön vapaamuotoisen potilasdokumentaation teho-osastoilla. Potilaan saapuessa teho-osastolle vastaanottava sairaanhoitaja kirjoittaa hänestä vapaamuotoisen tuloraportin. Tehohoidon aikana sairaanhoitajat ja fysioterapeutit kirjaavat vapaamuotoiset muistiinpanonsa kronologisesti niiden aiheen mukaisille potilastietojärjestelmän välilehdille. Tehohoitojakson lopuksi sairaanhoitaja kirjoittaa potilaasta uloskirjausraportin, johon hän tiivistää potilaan jatkohoidon kannalta tarpeelliset tiedot. Näiden asiakirjojen ensisijaisena tavoitteena on siirtää potilaan hoitoa koskevaa tietoa työntekijöiden välillä työvuorojen vaihtuessa tai potilaan siirtyessä osastolta toiselle.

Vapaamuotoisena tekstinä kirjatun tiedon suuri määrä rajoittaa sen jatkokäyttöä ja siten tiedonkulkua teho-osastoilla. Tutkimuksemme [26] mukaan erityisesti hoidon aikana kirjatut muistiinpanot ovat suuren kokonsa ja hajanaisen rakenteensa vuoksi hyödyntämätön voimavara: 516 suomalaisen pitkäaikaistehohoitopotilaan[1] asiakirjojen otoksessa[2] yhdestä potilaasta on kirjattu yhteensä keskimäärin 2100 sanaa hoidonaikaisia muistiinpanoja päivittäisen sanamäärän ollessa noin 200 sanaa. Enimmillään tätä tekstiä on kertynyt yhteensä jopa 13000 sanaa, eli noin 50 sivua. Tämän tekstimäärän lukeminen ja tiivistäminen vaativassa ja kiireisessä tehohoitoympäristössä on vaikea ja työläs tehtävä etenkin, kun tulo-, hoito- ja siirtovaiheen potilasdokumenttien tarkemman sisällönanalyysin[3] perusteella asiakirjojen rakenne ei ole järjestelmällinen. Esimerkiksi valinnaisesta ja vapaasti muokattavasta otsikoinnista huolimatta samasta aiheesta on kirjattu usean erityyppisen otsikon alle ja yhden otsikon alla on usein muistiinpanoja monista eri aiheista. Tämä tekee tiettyyn aiheeseen liittyvän tiedon etsimisestä haastavaa, mikä on sisällönanalyysissä nähtävissä tiedonkulun ongelmina hoidon aikana kirjatuista muistiinpanoista siirtotiedotteisiin.

3 Kieliteknologia ja potilasasiakirjat: etuja ja riskejä

Tutkimuksissamme [25] ja [24] kartoitetaan kieliteknologian sovelluskohteita teho-osastoilla ja pohditaan eettisiä näkökohtia, jotka tulee huomioida kehitettäessä ja käytettäessä kieliteknologiaa potilasasiakirjoille. Tutkimuksessamme [26] sovelluskohteet on yhdistetty tapaan, jolla tiedon tulisi kulkea teho-osaistoilla saumattomasti aina osastolle tulosta hoitoajan kautta siirtovaiheeseen. Tavoitteenamme on ollut luoda malli kieliteknologian kehittämiselle teho-osastojen tarpeisiin.

Tulostemme mukaan kieliteknologia voi tarjota uudenlaisia mahdollisuuksia helpottaa vapaamuotoisen potilasdokumentaation tuottamista ja hyödyntämistä. Esimerkkeinä kirjaamisen avustamisesta mainittakoon tekstin luettavuutta ja ymmärrettävyyttä parantavat oikoluku-, sanaston yksikäsitteistämis- (mm. yhtenäiset lyhenteet) ja tekstin jäsentelytoiminnot sekä tavat muistuttaa hoitajaa kirjoittamaan dokumentaatiota potilaan hoidon kannalta olennaisista asioista. Tekstin hyödyntämistä tukevia sovelluskohteita ovat vaikkapa potilaan tilan etenemistä

[1] Osastolla vähintään viisi päivää

[2] Kerätty asianmukaisin luvin aikaväliltä 1.1.2005–1.8.2006

[3] Kolme asiakirjaa kustakin vaiheesta

kuvaavien trendien, hälytyksien ja muistutuksien muodostaminen myös vapaamuotoisen potilasdokumentaation perusteella, aiheittainen tiedonhaku potilasasiakirjoista sekä uloskirjausraporttiluonnoksen teko automaattisesti kirjatusta potilasdokumentaatiosta tiivistämällä. Nämä mahdollisuudet tukevat myös hoitohenkilökunnan päätöksentekoa tavalla, joka samalla turvaa hoidon laatua, jatkuvuutta ja yksilöllisyyttä.

Tehohoidon kirjaamisprosessiin yhdistettynä kieliteknologian avulla yksittäisen potilaan hoitoa voidaan tukea paitsi häneen omiin asiakirjoihinsa tallennetun tiedon myös edeltävien potilaiden dokumenttien heijastaman käytännön tietämyksen ja tutkimuskirjallisuuden kuvaaman näytön avulla. Kieliteknologian ja teho-osaston tiedonkulun yhdistävässä mallissamme (ks. [26]) uuden potilaan siirtyessä tehohoitoon kieliteknologiaa käytetään yksilöllisen potilasprofiilin muodostamiseen hänen tuloraporttinsa ja edeltävien potilaiden dokumenttien vertailun perusteella. Tehohoitojakson aikana kieliteknologian avulla tuetaan hoitotyön vapaamuotoisten potilasasiakirjojen kirjoittamista ja henkilökunnan päätöksentekoa. Tehohoidon lopuksi kieliteknologialla avustetaan olennaisen tiedon tiivistämistä uloskirjausraporttiin esimerkiksi korostaen tekstistä tiettyyn aiheeseen liittyvät osiot sairaanhoitajan edelleen analysoitaviksi. Kieliteknologiakehityksen edetessä tätä lähestymistapaa voidaan edelleen jalostaa kohti kokonaan automatisoitua uloskirjausraporttiluonnoksen tekoa.

Kieliteknologian edut potilasdokumentaatiosovellusalueella voidaan saavuttaa vain silloin, kun on huomioitu kieliteknologian kehittämisen ja käytön eettiset riskit liittyen erityisesti luottamukselliseen datan käyttöön ja potilasturvallisuuteen [24]. Menetelmäkehityksessä riskit on huomioitava aina kehitystyössä tarvittavan potilasasiakirja-aineiston keruusta aineiston asianmukaiseen tuhoamiseen asti. Käytettäessä kieliteknologiakomponentteja terveydenhuollossa on muistettava kyseenalaistaa kieliteknologian tulokset, kuten saadut muistutukset tai niiden puute. Lisäksi on huolehdittava erityisen hyvin potilastietojen luottamuksellisuuden turvaamisesta väärinkäytön mahdollisuuksien nopeutuessa entistä tehokkaamman tiedonhaun sekä mahdollisesti syntyvän, ennestään tuntemattoman arkaluonteisen tiedon myötä.

4 Tiedonhaun tukeminen kieliteknologisin ratkaisuin

Luvussa tarkastellaan kahta tiedonhaun tukemiseen liittyvää osatutkimusta: tekstin aiheittaisen luokittelun ja relevanssin arvioinnin automatisointia (kohta 4.1) sekä tekstiaiheittaisten kappaleenjakojen tuottamista ja kappaleiden aiheiden nimeämistä automaattisesti (kohta 4.2). Vastaavantyyppisen tekstin rakenteistamisen on todettu nopeuttavan tiedonhakua potilasasiakirjoista [30].

4.1 Tekstin aiheittainen luokittelu ja relevanssin arviointi automaattisesti

Tutkimuksissamme [27] ja [6] tavoitteena on tukea tiedon hakua tehohoidon potilasasiakirjoista korostamalla kulloinkin tarkasteltavan aiheen kannalta olennaiset tekstikatkelmat (kuva 1). Tavoitetta lähestytään kokeilemalla koneoppimiseen perustuvien lähestymistapojen käyttökelpoisuutta potilasasiakirjojen tekstin automaattiseen luokitteluun ja relevanssiin perustuvaan järjestämiseen tarkastel-

tavien aiheiden ollessa hengitys, verenkierto ja kipu.

Tutkimuksissa käytetty aineisto koostuu keväällä 2001 suomalaisilta tehoosastoilta asianmukaisin luvin kerätystä anonymisoiduista hoitotyön vapaamuotoisesta potilasdokumentaatiosta. Tekstin sisällön asiantuntija on jakanut tutkimusaineiston yhden ajatuksen mittaisiksi katkelmiksi tuloksena 1363 tekstikatkelmaa. Tämän jälkeen kolme sairaanhoitajaa on merkinnyt toisistaan riippumattomasti tutkimusaineiston luokkiin Hengitys, Verenkierto ja Kipu liittyvät tekstikatkelmat. Koneoppimiseen perustuvien algoritmien opetusvaiheessa tämä aineisto on jaettu opetusjoukoksi (708 katkelmaa) ja testausjoukoksi (655 katkelmaa) siten, että alun perin samasta asiakirjasta peräisin olevia katkelmia ei ole jaettu sekä opetus- että testausjoukkoon.

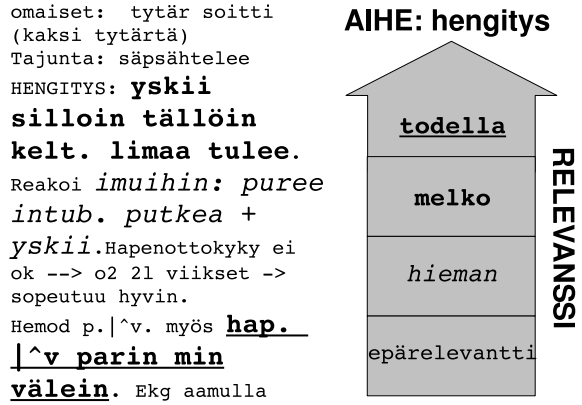
Tutkimuksessa [6] tarkastellaan kahta automaattisen luokittelijan toteuttamisen liittyvää asiaa. Aluksi selvitetään sairaanhoitajien yksimielisyyttä luokkien kannalta olennaisista tekstikatkelmista Cohenin κ-kertoimen [2] avulla. Cohenin κ-kertoimen arvo on keskimäärin 0,7 luokassa Hengitys, 0,9 luokassa Verenkierto ja 0,8 luokassa Kipu. Sairaanhoitajien mielipiteissä on siis joitakin eroavuuksia erityisesti luokkien Hengitys ja Kipu kohdalla.

Tutkimuksessa arvioidaan seuraavaksi koneoppimiseen perustuvan automaattisen luokittelijan suorituskykyä. Jokaiselle sairaanhoitaja–luokka-parille opetetaan oma koneensa ja luokitustuloksia verrataan toisiinsa käyttäen AUC-mittaa [5]. Luokittelutehtävä ratkaistaan tukivektorikoneille läheistä sukua olevan RLS-algoritmin (regularized least squares, ks. esim. [15]) avulla. Algoritmi perustuu tavoiteluokituksen ja luokittelijan tuottaman luokituksen vertailuun minimoiden neliövirhettä opetusaineistolla sekä luokittelijan yleistettävyyttä opetusaineistosta kontrolloivan parametrin käyttöön. AUC-mitan arvot ovat keskimäärin 0,9 luokissa Hengitys ja Verenkierto sekä keskimäärin 0,7 luokassa Kipu, kun luokittelijan tuottamia tuloksia verrataan saman sairaanhoitajan mielipiteisiin, jonka merkitsemää aineistoa on käytetty koneen opettamisessa. Toisessa arvioinnissa luokittelijan tuloksia verrataan muiden sairaanhoitajien kuin opetusdatan muodostaneen sairaanhoitajan merkintöihin. Tällöin havaitaan, että luokassa Verenkierto muiden kuin koneen suorituskyvyn arvioinnissa käytetyn aineiston merkinneen sairaanhoitajan mielipiteisiin perustuvat luokittelijat toimivat yhtä hyvin kuin luokittelija, joka on opetettu ja testattu saman sairaanhoitajan mielipiteisiin perustuvilla aineistoilla.

Tutkimus [27] jatkaa edellä mainittua hyödyntäen sairaanhoitajien mielipideeroja luokan kannalta olennaisista tekstikatkelmista. Tavoitteena on järjestää tekstikatkelmat sen mukaan, miten relevanttina niitä voidaan pitää kulloinkin tarkasteltavan aiheen kannalta. Tällöin suurimman relevanssin omaavia katkelmia voisi käyttää esimerkiksi olennaisen tiedon tiivistämiseen uloskirjausraporttiin ja matalan relevanssin omaavia katkelmia vaikkapa hälytystoimintoihin.

Koneoppimiseen perustuvan algoritmin opettamisessa ja testauksessa käytettävä aineisto muodostetaan laskemalla jokaiselle tekstikatkelmalle sen tiettyyn luokkaan kuuluvaksi merkinneiden hoitajien lukumäärä. Tulosta tulkitaan siten, että relevanssin kolme omaavat tekstikatkelmat ovat tietyn luokan kannalta olennaisimpia ja katkelman yhteys luokkaan heikkenee relevanssin pienentyessä kohti nollaa.

Koneen tuottamia relevanssiastei-



Kuva 1: Esimerkki tehostetusta tiedonhausta potilasasiakirjoista hyödyntäen tekstin korostamista relevanssin mukaisesti.

ta verrataan manuaalisesti aikaansaatuihin pisteisiin käyttäen Kendallin τ_b-mittaa [8]. Relevanssin arviointitehtävä ratkaistaan tarkasteltavaan ongelmaan sovitetulla RLS-algoritmilla. Järjestyksien välinen assosiaatio on voimakkain luokissa Hengitys ($\tau_b \approx 0{,}6$) ja Verenkierto ($\tau_b \approx 0{,}7$). Luokassa Kipu järjestyksien välinen assosiaatio on jonkin verran heikompi ($\tau_b \approx 0{,}4$). Heikompi assosiaatio luokassa Kipu voidaan katsoa johtuvan tähän luokkaan liittyvän dokumentaation erilaisesta luonteesta sekä muita luokkia vähäisemmästä esiintymismäärästä. Näin ollen tarkasteltaviin luokkiin liittyvän tekstin määrä on huomioitava opetusjoukon riittävänä kokona ja luonne kieliteknologiasovelluksen yksityiskohdissa.

4.2 Kappaleenjaot ja kappaleiden aiheet automaattisesti

Tutkimuksemme [28] ja [4] jatkavat tiedonhakuongelman ratkaisemista kieliteknologian avulla (kuva 2): tavoitteena on jakaa tehohoidon potilasdokumentaation vapaatekstiosiot automaattisesti aiheidensa mukaisiin kappaleisiin[4] ja nimetä ne aiheittain.

Tutkimusaineistona käytetään luvussa 3 mainittujen 516 potilaan asiakirjojen hoidon aikaisia tekstejä. Ongelmanasettelussa kappaleenjaot määrittäviksi hakuaiheiksi on valittu aineiston selvästi yleisimmät aiheet: hengitys, hemodynamiikka, tajunta, omaiset ja diureesi. Koneoppimiseen perustuvan kieliteknologian opettamiseksi ja testaamiseksi 135 hoitoajankohdan mukaan ensimmäisen potilaan asiakirjoista on poimittu satunnaisten kolmen hoitotyövuoron kirjaukset ja näihin on merkitty hakuaiheisiin liittyvät tekstikatkelmat manuaalista sisällönanalyysiä käyttäen. Aiheisiin liittymättömät tekstikatkelmat on jätetty aiheiden ulkopuoliseen luokkaan Muuta kuuluviksi. Yleensä hoitotyövuoron kirjauksessa käsitellään lähes kaikkia viittä aihetta keskimääräisen kappalepituuden ollessa 18 sanaa.

Erona edeltäviin tutkimuksiin [6] ja [27] on, että tekstikatkelman sallitaan kuuluvan vain yhteen aiheeseen kerrallaan. Valinta perustuu aineistossa yleisesti käytettyyn kirjausrakenteeseen, jos-

[4] Ongelma tunnetaan nimellä aihesegmentointi.

sa yllä mainitut viisi aihetta ovat käytössä toistensa poissulkevina otsikoina, vaikkakaan otsikot eivät olleet yhdenmukaisia. Lisäksi muutoksella haetaan automatisoitua ratkaisua tutkimuksissa [6] ja [27] sisällön asiantuntijan käsin tekemään tekstin jakoon kappaleiksi yhdistäen kappaleenjako- ja aiheiden nimeämistehtävät yhdeksi ongelmaksi.

Vertailtavia menetelmiä on yhteensä kolme (kuva 2). Vertailun lähtökohtana on kirjausrakennetta muistuttava aihehakualgoritmi. Sen toimintaperiaatteena on hakea tekstistä viittä mielenkiinnon kohteena olevaa aihetta, siis sanoja *hengitys*, *hemodynamiikka*, *tajunta*, *omainen* ja *diureesi*. Algoritmi merkitsee kullekin tekstin sanalle aiheen ottaen aina hoitotyövuoron alussa alkuarvokseen luokan Muuta ja tämän jälkeen merkiten vuoron loppuun asti kunkin sanan viimeksi esiintyneeseen hakuaiheeseen.

Ohjattuun oppimiseen perustuvaa ensimmäisen asteen kätkettyä Markovin mallia (KMM, ks. esim. [20]) käytetään ensimmäisenä ongelman varsinaisena ratkaisutapana. Mallin opettamiseksi ja testaamiseksi käsin analysoitu aineisto on jaettu suunnilleen puoliksi siten, että saman potilaan asiakirjasta peräisin olevia vuoroja ei ole jaettu sekä opetus- että testausjoukkoon[5] [28].

Ratkaisutavoista seuraava [4] kehittää ensimmäistä muodostamalla KMM:a muistuttavan graafisen mallin, joka mahdollistaa hakuaiheiden asettamisen, mutta ei vaadi manuaalisesti analysoitua opetusaineistoa. Toisin sanoen tavoitteena on toteuttaa Internetin hakukoneita muistuttava sovellus, jolle käyttäjä antaa häntä kiinnostavat aiheet hakusanoina (esim. *hengitys*, *tajunta*, *diureesi*) lopputuloksen ollessa potilasasiakirja, josta aiheita vastaavat osiot eroteltu esimerkiksi aiheittain omalla väärillään. Mallin muodostamisessa käytetään apuna Schützen Word Space -versiota latentista semanttisesta analyysistä [21] sekä aiheittaisia kappalepituuksia kontrolloivaa parametria.

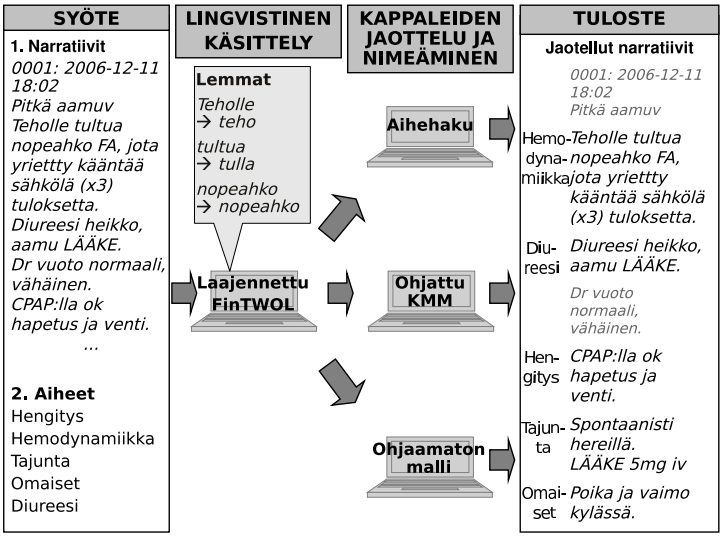
Tekstiä on yhtenäistetty ennen vertailuja suomenkielen morfologiseen analyysiin tarkoitetulla FinTWOL-ohjelmalla [9] korvaamalla jokainen ohjelman tunnistama sana sitä vastaavan tulosteen ensimmäisellä lemmalla (kuva 2). Siis esimerkiksi sanat *teholle* ja *teholla* korvataan perusmuodolla *teho*. Lingvistisen käsittelyn kattavuuden parantamiseksi ohjelman sanastoa on laajennettu noin 3500 kliinisellä sovellusalan sanalla.

Suorituskyvyn vertailu toteutetaan vertailemalla menetelmien tuottamia aiheita manuaalisessa sisällönanalyysissä annettuihin ja laskemalla sanoittainen tarkkuuden[6] keskiarvo koko testausjoukossa. Vertailun lähtökohtana olleen aihehakualgoritmin tarkkuus on noin 67%. Ohjattuun oppimiseen perustuva KMM päihittää tämän tuloksen jo 2000 sanan (noin 25 hoitotyövuoron kirjaukset) opetusaineistolla. Koko opetusjoukko (noin 16000 sanaa) hyödyntäen KMM:n tarkkuus on 83%. Ilman lingvististä käsittelyä FinTWOLia käyttäen vastaava luku on 81%. Ohjaamattoman mallin suorityskyky, 75%, on huomattavasti parempi kuin yksinkertaisen ohjaamattoman aihehakualgoritmin, mutta odotetusti heikompi kuin paljon enemmän ongelmasta tietoa saavan ohjatun KMM:n. Ohjattu KMM toimii ohjaamatonta mallia paremmin, kun opetukseen käytetään vähintään 3600 sanaa.

Tutkimuksien lopputuloksena on, että tiedonhakuongelman tulee määrittää se, valitaanko ohjattu vai ohjaamaton mene-

[5] Opetus- ja testausvaiheiden yksityiskohtien kuvaus tutkimuksessa [28]

[6] Oikein luokiteltujen sanojen suhde tekstin kokonaissanamäärään



Kuva 2: Tutkimusasetelma kappaleenjakojen tuottamisen ja kappaleiden aiheiden nimeämisen automatisointia tarkastelevissa tutkimuksissa [28] ja [4].

telmä: ohjattu KMM on ensisijainen vaihtoehto, mikäli hakuaiheet ovat ennalta tiedossa ja resurssit opetusaineiston muodostamiseen ovat olemassa. Ohjaamaton vaihtoehto tulisi valita, kun halutaan etsiä vapaasti valittavia aiheita ja opetusaineistoa ei ole.

5 Jatkotutkimuksesta

Vuosina 2005–2008 saamamme tulokset rohkaisevat jatkotutkimusta: ensinnäkin olemme havainneet, että kieliteknologian menetelmät tarjoavat lukuisia mahdollisuuksia olemassa olevien potilastietojärjestelmien ominaisuuksien kehittämiseen. Toiseksi olemme kyenneet käyttämään kieliteknologiaa menestyksekkäästi tiedonhakuongelmien ratkaisuun vapaamuotoisesta potilasdokumentaatiosta.

Jatkamme potilasdokumentaation jatkohyödyntämiseen soveltuvan kieliteknologian kehittämistä tarkastelemalla myös muita kirjaamisen osa-alueita ja muita tekstin hyödynnettävyyteen liittyviä ongelmia. Lisäksi laajennamme tutkimusta teho-osastoilta muillekin terveydenhuollon osa-alueille. Suunnitteilla on myös aidossa hoitoympäristössä menetelmiä testaava pilottitutkimus.

Kiitokset

Lämmin kiitos Tietojenkäsittelytieteen Seura ry:n hallitukselle työni saamasta tunnustuksesta Tietojenkäsittelytieteen päivillä 2007 sekä Tekesille väitöstutkimukseni rahoittamisesta osana Louhi-hanketta (rahoituspäätökset 40435/05 ja 40020/07). Kiitän yhteistyöstä myös kaikkia muita hankkeen osapuolia, erityisesti Sari Ahosta ja Simo Vihjasta Lingsoft Oy:stä, väitöstutkimukseni ohjaajia Eija Helena Karstenia Åbo Akademista ja Tapio Salakoskea Turun yliopistolta sekä kaikkia artikkeleihini osallistuneita tutkijoita.

Viitteet

[1] Ambrosius, Huittinen, V-M., Kari, A., Leino-Kilpi, H., Niinikoski, J., Ohtonen, M., Rauhala, V., Tammisto, T. ja Takkunen, O. (1997). Suomen tehohoitoyhdistyksen eettiset ohjeet, *Tehohoitolehti*, osa 15, nro 2, 165–172.

[2] Cohen, J. (1960). A coefficient of agreement for nominal scales, *Educational and psychological measurement*, osa 20, nro 2, 37–46.

[3] Friedlin, J. ja McDonald, C.J. (2006). Using a natural language processing system to extract and code family history data from admission reports, *AMIA ... Annual Symposium proceedings / AMIA Symposium*, 925.

[4] Ginter, F., Suominen, H., Pyysalo, S. ja Salakoski, T. (2008). Combining hidden Markov models and latent semantic analysis for topic segmentation and labeling: method and clinical application, teoksessa: T. Salakoski, D. Rebholz-Schuhmann, S. Pyysalo (toim.) *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine* (*SMBM 2008*), TUCS General Publications, osa 51, 37–44.

[5] Hanley, J.A. ja McNeil, B.J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, osa 143, nro 1, 29–36.

[6] Hiissa, M., Pahikkala, T., Suominen, H., Lehtikunnas, T., Back, B., Karsten, H., Salanterä, S. ja Salakoski, T. (2007). Towards automated classification of intensive care nursing narratives, *International journal of medical informatics*, osa 76, nro S3, S362–S368.

[7] Hripcsak, G., Knirsch, C., Zhou, L., Wilcox, A. ja Melton, G.B. (2007). Using discordance to improve classification in narrative clinical databases: an application to community-acquired pneumonia, *Computers in biology and medicine*, osa 37, nro 3, 296–304.

[8] Kendall, M.G. (1970). *Rank correlation methods*. 4. painos, Griffin, Lontoo.

[9] Koskenniemi, K. (1983). Two-level model for morphological analysis, teoksessa: *Proceedings of International Joint Conference on Artificial Intelligence 1983 (IJCAI-83)*, Morgan Kaufmann, 683–685.

[10] Lauri, S. ja Salanterä, S. (2002). Developing an instrument to measure and describe clinical decision making in different nursing fields, *Journal of professional nursing*, osa 18, nro 2, 93–100.

[11] Louhi (2008). http://www.med.utu.fi/hoitotiede/tutkimus/tutkimusprojektit/louhi/ (haettu 06.10.2008).

[12] Lovis, C., Baud, R.H. ja Planche, P. (2000). Power of expression in the electronic patient record: structured data or narrative text?, *International journal of medical informatics*, osa 58–59, 101–110.

[13] Mendonça, E.A., Haas, J., Shagina, L., Larson, E. ja Friedman, C. (2005). Extracting information on pneumonia in infants using natural language processing of radiology reports, *Journal of biomedical informatics*, osa 38, nro 4, 314–321.

[14] Meystre, S. ja Haug, P. (2006). Natural language processing to extract medical problems from electronic clinical documents: Performance evaluation, *Journal of biomedical informatics*, osa 39, nro 6, 589–599.

[15] Pahikkala, T. (2008). *New kernel functions and learning methods for text and data mining*. TUCS Dissertations, nro 103, Turku Centre for Computer Science, Turku.

[16] Pakhomov, S.V., Buntrock, J.D. ja Chute, C.G. (2006). Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques, *Journal of the American Medical Informatics Association*, osa 13, nro 5, 516–525.

[17] Pakhomov, S.S., Hemingway, H., Weston, S.A., Jacobsen, S.J., Rodeheffer, R. ja Roger, V.L. (2007). Epidemiology of angina pectoris: role of natural language processing of the medical record, *American heart journal*, osa 153, nro 4, 666–673.

[18] Pakhomov, S., Weston, S.A., Jacobsen, S.J., Chute, C.G., Meverden, R. ja Roger, V.L. (2007). Electronic medical records for clinical research: application to the identification of heart failure, *The American journal of managed care*, osa 13, nro 6, 281–288.

[19] Pentz, J.F., Wilcox, A.B. ja Hurdle, J. (2007). Automated identification of adverse events related to central venous catheters, *Journal of biomedical informatics*, osa 40, nro 2, 174–182.

[20] Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition *Proceedings of the IEEE*, osa 77, nro 2, 257–286.

[21] Schütze, H. (1998). Automatic word sense discrimination, *Computational Linguistics*, osa 24, nro 1, 97–123.

[22] Shah, A.D. ja Martinez, C. (2006). An algorithm to derive a numerical daily dose from unstructured text dosage instructions, *Pharmacoepidemiology and drug safety*, osa 15, nro 3, 161–166.

[23] Suominen, H. (2007). Kieliteknologian menetelmien soveltaminen potilasdokumentaation hyödyntämiseen, teoksessa M. Koskinen, E. Jauhiainen (toim.) *Tietojenkäsittelytieteen päivät 2007*, Tietojenkäsittelytieteiden julkaisuja, Tutkimuksia TU-25, Jyväskylän yliopistopaino, Jyväskylä, 46–50.

[24] Suominen, H., Lehtikunnas, T., Back, B., Karsten, H., Salakoski, T. ja Salanterä, S. (2007). Applying language technology to nursing documents: pros and cons with a focus on ethics, *International journal of medical informatics*, osa 76, nro S2, S293–S301.

[25] Suominen, H. J., Lehtikunnas, T., Hiissa, M., Back, B., Karsten, H., Salakoski, T. ja Salanterä, S. (2005). Natural language processing for nursing documentation, teoksessa: J. M. Fonseca (toim.) *Proceedings of the 2nd International Conference on Computational Intelligence in Medicine and Healthcare*, 147–154.

[26] Suominen, H., Lundgrén-Laine, H., Perttilä, J., Salakoski, T. ja Salanterä, S. (2008). Tehohoidon elektroniset potilasasiakirjat — hyödyntämätön voimavara, teoksessa *Valtakunnalliset Lääkäripäivät 2008*.

[27] Suominen, H., Pahikkala, T., Hiissa, M., Lehtikunnas, T., Back, B., Karsten, H., Salanterä, S. ja Salakoski, T. (2006). Relevance ranking of intensive care nursing narratives, *Lecture Notes in Computer Science*, osa 4251, 720-727.

[28] Suominen, H., Pyysalo, S., Ginter, F. ja Salakoski, T. (2008). Automated text segmentation and topic labeling of clinical narratives, teoksessa: B. Back, H. Karsten, T. Salakoski, S. Salanterä, H. Suominen (toim.) *Proceedings of the First Conference on Text and Data Mining of Clinical Documents* (*Louhi'08*), TUCS General Publications, osa 52, 99–103.

[29] Tange, H.J., Hasman, A., de Vries Robbe, P.F. ja Schouten, H.C. (1997). Medical narratives in electronic medical records, *International journal of medical informatics*, osa 46, nro 1, 7–29.

[30] Tange, H.J., Schouten, H.C., Kester, A.D.M, ja Hasman, A. (1998). The granularity of medical narratives and its effect on the speed and completeness of information retrieval, *Journal of the American Medical Informatics Association: JAMIA*, osa 5, nro 6, 571–582.

[31] Turchin, A., Kolatkar, N.S., Grant, R.W., Makhni, E.C., Pendergrass, M.L. ja Einbinder, J.S. (2006). Using regular expression to abstract blood pressure and treatment intensification information from the text of physician notes, *Journal of the American Medical Informatics Association: JAMIA*, osa 13, nro 6, 691–695.

[32] Zeng, Q.T., Goryachev, S., Weiss, S., Sordo, M., Murphy, S.N. ja Lazarus, R. (2006). Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system, *BMC medical informatics and decision making*, osa 6, 30.

[33] Zhou, L., Tao, Y., Cimino, J.J., Chen, E.S., Liu, H., Lussier, Y.A., Hripcsak, G. ja Friedman, C. (2006). Terminology model discovery using natural language processing and visualization techniques, *Journal of biomedical informatics*, osa 39, nro 6, 626–636.