Tapaustutkimus: semanttinen linkitys ja tiedon löydettävyys puolustusvoimien normitietokannassa

Mika Wahlroos, Matias Frosterus ja Eero Hyvönen

Semanttisen laskennan tutkimusryhmä (SeCo)
Aalto-yliopisto, Perustieteiden korkeakoulu, Mediatekniikan laitos, ja
Helsingin yliopisto, Tietojenkäsittelytieteen laitos
http://www.seco.tkk.fi/
etunimi.sukunimi@aalto.fi

Tiivistelmä Asiakirjahallinnossa käytetään yleisesti metatietoja asiakirjallisen tiedon kuvailuun. Semanttisessa webissä metatieto pyritään yhdistämään muuhun vastaavaan tietoon käyttämällä standardoituja metatiedon esitystapoja ja ontologiatekniikoita, jolloin metatiedossa esiintyvien käsitteiden väliset semanttiset yhteydet saadaan ilmaistua koneellisesti luettavassa muodossa. Olemme selvittäneet puolustusvoimien normitietokannan metatietojen yhdistämistä ontologiaan, metadatan puoliautomaattista eristämistä normien tekstistä sekä näin syntyneen semanttisesti rikastetun metatiedon hyödyntämistä asiakirjallisen tiedon hallinnassa.

1 Johdanto

Asiakirjahallinnossa tiedon kuvailuun käytetään useita erilaisia metatietoja. Asiakirjallisen tiedon käsittelyssä yleisesti käytettäviä metatietoja ovat otsikoiden ja yhteenvetojen lisäksi mm. asiasanat sekä asian käsittelyyn liittyvät henkilöt, organisaatiot ja päivämäärät. Pyrkimyksenä on parantaa tiedon hallittavuutta tai löydettävyyttä. Asiakirjaan liittyviä metatietoja kutsutaan myös asiakirjan *annotaatioiksi* ja metatietojen lisäämistä vastaavasti tiedon *annotoinniksi*.

Semanttisessa webissä [1] metatiedot esitetään RDF-standardin[1] (Resource Description Framework) mukaisessa muodossa, jossa kukin yksittäinen käsite tai muu instanssi esitellään erillisenä *resurssina*. Kullakin resurssilla on oma yksikäsitteinen tunnisteensa, jonka avulla siihen voidaan viitata muusta tiedosta käsin, jolloin toisiinsa viittaavat resurssit muodostavat merkitysten verkon. Semanttisessa webissä käytettävä tunnistetyyppi on verkko-osoitteen muotoinen URI (Uniform Resource Identifier). *Ontologioissa* puolestaan on määritelty käsitteiden välisiä merkityssuhteita ja muita keskinäisiä yhteyksiä koneellisesti luettavassa muodossa [3]. Tämä mahdollistaa merkityssuhteiden hyödyntämisen esimerkiksi tiedon löydettävyyden parantamiseksi semanttisissa haku- ja suosittelujärjestelmissä.

FinnONTO 2.0 -hankkeessa[2] on tutkittu semanttisen webin teknologioiden käyttöä puolustusvoimien asianhallintajärjestelmissä. Tavoitteena oli selvittää uusien teknolo-

[1] http://www.w3.org/RDF/
[2] http://www.seco.tkk.fi/projects/finnonto/

gioiden mahdollisuuksia asiakirjallisen tiedon indeksoinnissa ja haussa tapaustutkimuksella puolustusvoimien normitietokannasta. Tutkimuksessa käytetty suomenkielisisten julkisten normien metatiedoista koostuva aineisto muunnettiin semanttisesti yhdistettävään muotoon minkä jälkeen luotiin hakukäyttöliittymä, jonka avulla aineistosta voidaan hakea tietoa monista eri näkökulmista. Lisäksi tutkittiin menetelmiä metatiedon automaattiseksi lisäämiseksi tekstiaineiston perusteella.

Tämän artikkelin luvussa 2 esitellään yleisesti ontologioita ja niiden käyttöä sekä normien metatietojen semanttista yhdistämistä ontologiaan. Luvussa 3 käsitellään vapaan asiasanoituksen suhdetta ontologioihin ja kontrolloituihin asiasanastoihin, ja luvussa 4 selvitetään käytettävissä olevan metatiedon määrän kasvattamista automaattisten annotointimenetelmien avulla. Luvussa 5 pohditaan tulevia kehityssuuntia, ja luvussa 6 esitetään yhteenveto ja johtopäätökset.

2 Metatietojen semanttinen yhdistäminen

Tässä luvussa esitellään yleisesti ontologioita ja niiden käyttöä sekä kuvataan prosessi puolustusvoimien normien metatietojen muuntamiseksi semanttisesti yhdistettävään muotoon.

2.1 Ontologiat

Ontologialla tarkoitetaan muodollista tietämyksen esitystapaa, jossa tietyn toimialan piiriin kuuluvat käsitteet ja niiden väliset keskinäiset suhteet ilmaistaan koneellisesti käsiteltävässä muodossa. Yksinkertaisimmillaan ontologia voi olla käsitehierarkia, jossa eri käsitteitä on määritelty toistensa ylä- tai alakäsitteiksi, mutta myös muunlaisia yhteyksiä voidaan esittää. Ontologisilla suhteilla voidaan esimerkiksi ilmaista, että kaksi eri sanaa ovat toistensa synonyymeja, tai eritellä toisistaan monitulkintaisen sanan eri merkitykset. Vastaavasti esimerkiksi paikkaontologioita voidaan käyttää ilmaisemaan koneellisesti käsiteltävässä muodossa erilaisten sijaintien tai alueiden keskinäisiä suhteita. Koska tietokoneelta puuttuvat tietämys ja taustatieto, joita käsitteiden merkitysten ymmärtäminen edellyttää, käsitteiden keskinäisten suhteiden täsmällinen määrittely tekee ontologisista sanastoista koneellisen käsittelyn kannalta perinteisiä asiasanastoja hyödyllisempiä [6].

Tässä tutkimuksessa ontologiana käytettiin Puolustushallinnon ontologiaa PUHO:a[3], joka on kehitetty FinnONTO-hankkeessa aiemmin Puolustushallinnon asiasanaston pohjalta.

2.2 Metatietojen yhdistäminen ontologiaan

Järjestelmien yhteentoimivuuden vuoksi metatiedot on esitettävä standardoidussa muodossa. Eräs yleisessä käytössä oleva standardisanasto metatiedon kuvaamiseen on Dublin Core[4], jossa on määritelty joukko dokumenttien metatietoelementtejä mm. asiasanojen,

[3] http://onki.fi/fi/browser/overview/puho
[4] http://dublincore.org/

kuvauksen, asiakirjan laatijan ja erilaisten päivämäärätietojen ilmaisemiseen. Dublin Coren pohjalta on kehitetty myös mm. arkistolaitoksen SÄHKE2-standardi ja -määräykset[5]. Standardimuotoinen esitystapa on tärkeää myös tietojen semanttisen yhdistämisen kannalta.

Tutkimushanketta varten saimme otoksen aineistosta, joka koostui normitietokannan julkisista asiakirjoista ja niiden metatiedoista. Metatiedot siirrettiin tutkimusryhmälle SÄHKE2-metatietoskeeman mukaisessa XML-muodossa. Metatietojen muunnos XML-muodosta RDF-mallin mukaiseen muotoon tehtiin seuraavalla nelivaiheisella prosessilla:

1. Ontologian (PUHO) lemmaus
2. Metatietojen esikäsittely
3. RDF-metatietomallin luonti
   - Asiasanojen lemmaus
   - Metatietoelementtien muuntaminen sopivaan muotoon
4. RDF-mallin jälkikäsittely

Tekstimuotoisen tiedon *lemmauksella* tarkoitetaan sanojen muuntamista perusmuotoonsa. Kun ontologiassa määriteltyjen käsitteiden nimiöt ja normien metatiedoissa esiintyvät asiasanat on muunnettu perusmuotoonsa, kullekin asiasanastossa määritellylle asiasanalle voidaan löytää vastaava ontologinen käsite suoraan vertaamalla asiasanoja ontologisten käsitteiden nimiöihin, vaikka ne esiintyisivät metatiedoissa ja ontologiassa alunperin eri taivutusmuodoissa. Tämä on erityisen tärkeää agglutinatiivisten kielten kuten suomen tapauksessa. Asiasanojen ja ontologisten käsitteiden lemmaukseen käytettiin Connexor Oy:n kaupallista FDG-jäsennintä [12].

Edellä kuvattu prosessi ei kuitenkaan ratkaise moniselitteisyyden ongelmaa, jossa usealla merkitykseltään erillisellä käsitteellä voi olla sama perusmuotoistettu nimiö. Esimerkiksi hävittäjä voi tarkoittaa sekä laivaa että lentokonetta. Koska aineisto koostuu yksittäisistä asiasanoista ilman teksti- tai asiayhteyttä, tämä ongelma sivuutettiin valitsemalla ensimmäinen nimiön perusteella täsmäävä käsite.

Asiasanat, joita vastaava käsite löydettiin tällä tavoin ontologiasta, korvattiin RDF-mallissa URI-viittauksilla kyseisiin ontologiassa määriteltyihin käsitteisiin. Metatiedoissa on käytössä myös vapaita asiasanoja, jotka normin laatija on valinnut asiasanaston ulkopuolelta. Tällaisille asiasanoille ei välttämättä löydy suoraan vastaavaa käsitettä ontologiasta. Vapaat asiasanat lisättiin RDF-malliin erillisinä ontologian ulkopuolisina käsitteinä. Luvussa 3 käsitellään tarkemmin vapaiden asiasanojen käsittelyä ja suhdetta ontologisiin käsitteisiin.

Asiasanojen lisäksi myös valikoituja muita metatietoelementtejä muunnettiin sopivampaan muotoon ja lisättiin RDF-malliin käyttäen tarkoitukseltaan sopivia elementtejä esimerkiksi Dublin Coresta. Esimerkkejä tällaisista muunnoksista ovat mm. muuntaminen kansainvälisen ISO 8601 -standardin mukaiseen muotoon ja otsikkotiedon tallettaminen RDF-malliin käyttäen Dublin Coren vastaavaa title-elementtiä. Tässä artikkelissa keskitytään kuitenkin erityisesti asiasanojen käsittelyyn.

[5] http://www.arkisto.fi/fi/saehke2-maeaeraeys/

2.3 Metatiedon muokkaus ja haku

Edelläkuvatulla tavalla RDF-mallin mukaiseen muotoon muunnettu metadata tuotiin tämän jälkeen SAHA3-metatietoeditoriin [7], jonka yhteyteen on integroitu HAKO-moninäkymähakukone. SAHA3 mahdollistaa metatietojen muokkaamisen, esimerkiksi useiden samaa tarkoittavien vapaiden asiasanojen merkitsemisen synonyymeiksi. Hakukäyttöliittymä puolestaan mahdollistaa yksinkertaisten hakujen tekemisen aineistosta ja hakutulosten rajaamisen eri metatietokenttien perusteella [5]. Esimerkiksi suuren määrän tuloksia tuottaneen tekstihaun tulokset voi rajata esittämään vain tiettyyn asiakirjalajiin kuuluvat normit.

3 Vapaiden asiasanojen yhdistäminen ontologiaan

Vapaavalintaisia asiasanoja voidaan käyttää asiakirjallisen tiedon annotoinnissa asiasanaston ohella, jos asiasanaston ennaltamääritellyt termit eivät asiakirjan laatijan tai asiasanoittajan mielestä riitä kuvaamaan asiakirjan sisältöä riittävästi. Ontologian hyödyt jäävät kuitenkin vapaiden asiasanojen osalta hyödyntämättä, koska tällaiset asiasanat jäävät roikkumaan ontologian määrittelemän käsiteverkoston ja semanttisten yhteyksien ulkopuolelle.

Ontologian hyödyt saataisiin toki koskemaan myös vapaita asiasanoja laajentamalla ontologia kattamaan myös uudet käsitteet sitä mukaa kuin niitä syntyy vapaiden asiasanojen käytön myötä. Ontologian kehitys ja käsitteiden keskinäisten suhteiden asianmukainen ja täsmällinen määrittely on kuitenkin työlästä ja aikaavievää ja vaatii erityisosaamista. Eräs tapa tuoda semanttisten yhteyksien hyötyjä vapaille asiasanoille on määritellä vapaille asiasanoille yhteyksiä ontologisiin käsitteisiin itse aineistossa ontologian muokkaamisen sijaan [2]. Yhteyden luominen ontologiseen hierarkiaan määrittelee vapaan asiasanan kontekstin ja siten merkityksen täsmällisesti, jotta muutkin asiakirjojen laatijat voivat käyttää samaa termiä johdonmukaisesti. Käytännön asianhallintajärjestelmässä tällaisia kevyitä semanttisia yhteyksiä voitaisiin määritellä vapaille asiasanoille joko tarpeen mukaan jälkeenpäin tai osana asiasanoitusprosessia, jossa uusia vapaita asiasanoja syntyy.

Osana tutkimustyötä etsimme aineistosta normien metatiedoissa yleisimmin esiintyvät vapaat asiasanat ja lisäsimme niille viittauksia ontologisiin käsitteisiin. Etenkin yleisesti käytettyjen vapaiden asiasanojen osalta tietoja termien käytöstä voidaan mahdollisesti hyödyntää asiasanaston ja ontologian kehitystyössä. Taulukossa 1 on listattu yleisimpiä tutkimusaineistossa esiintyneitä vapaita asiasanoja ja niiden esiintymiskertoja.

Joidenkin vapaiden asiasanojen lukuisat esiintymiskerrat selittyvät tiettyjä aiheita, esimerkiksi turvallisuusasioita, käsitteleviä normeja ja niiden asiasanoitusta koskevilla määräyksillä. Muista asiasanoista jotkin esiintyvät vain yhden tai kahden organisaation tai muutaman henkilön käyttäminä, kun taas osaa on käytetty laaja-alaisemmin. Etenkin yleisimmät tai laajan käyttäjäpohjan käytössä olevat asiasanat ovat varteenotettavia huomioitaviksi ontologian jatkokehityksessä.

Lisäksi aineistosta löydettiin jonkin verran samaa tai lähes samaa tarkoittavia vapaita asiasanoja. Tällaisten termien osalta ontologisten yhteyksien määrittelyn eräs mahdollinen hyöty on asiasanoituksen yhdenmukaistaminen myös silloin, kun käsitettä tai

Asiasana	Esiintymiskerrat
palvelusturvallisuus	91
henkilöstöturvallisuus	44
tekninen hyväksyntä	34
henkilöturvallisuus	25
käyttöön hyväksyntä	16
räjähdeturvallisuus	16
toimintaohje	13
kuljetusturvallisuus	12
osaamisen kehittäminen	12
palkatun henkilöstön koulutus	12
kirkollinen työ	12
hallinnolliset asiat	10
RSRAKH	10
asevelvollisuusalan normit	8
varastointiturvallisuus	8

Taulukko 1. Yleisimmät vapaat asiasanat normien julkisissa metatiedoissa

sen kaikkia eri esiintymismuotoja ei haluta tai voida välittömästi lisätä varsinaiseen ontologiaan.

4 Automaattinen asiasanoitus

Tiedon annotointi eli metatietojen määrittäminen on usein työlästä ja aikaavievää. Jotkin metatiedot, esimerkiksi asiakirjan luontiajankohta, syntyvät luonnollisesti asiakirjallisen tiedon laatimisen yhteydessä. Hyvien, tiedon hallittavuutta tai löydettävyyttä parantavien asiasanojen valinta on kuitenkin vaikeaa. Metatiedon tarpeellisuus ja merkitys voivat jäädä epäselviksi tiedon laatijalle, mikä voi heikentää laatijan motivaatiota ja hankaloittaa entisestään tarkoituksenmukaisten asiasanojen valitsemista. Asiasanoitus voi tällöin jäädä vähäiseksi, puuttua kokonaan tai jäädä heikkolaatuiseksi [4].

Osittaiseksi ratkaisuksi metatiedon vähäisyyteen tai täydelliseen puuttumiseen on pyritty kehittämään erilaisia automaattisia tai puoliautomaattisia annotointimenetelmiä. Puoliautomaattisen asiasanoituksen tarkoituksena on ehdottaa automaattisesti asiasanoja tekstin sisällön perusteella, ja tiedon laatija voi harkintansa mukaan valita automaattisista ehdotuksista sopivat ja lisätä muita asiasanoja perinteiseen tapaan. Tarkoituksena ei tällöin ole korvata ihmisen asiantuntemusta teknisin keinoin vaan ensisijaisesti avustaa asiantuntijaa tehtävässään.

4.1 Normien puoliautomaattinen asiasanoitus

Osana tutkimustapausta olemme selvittäneet puoliautomaattisen asiasanoituksen soveltuvuutta käytettäväksi normien asiasanoituksen tukena. Automaattiseen asiasanoitukseen käytimme ARPA-rajapintaa, joka mahdollistaa erilaisten annotointimoottoreiden käytön yhteisen ohjelmallisen rajapinnan kautta [8]. Arviointia varten ARPA asetettiin

käyttämään Maui-annotointimoottoria [9], jonka on aiemmin todettu tuottavan hyviä tuloksia myös suomenkielisellä aineistolla [11]. Normitietokannan aineistolla saadut tulokset ovat alustavia, ja niitä on tarkoitus arvioida tarkemmin lähitulevaisuudessa. Esittelemme kuitenkin tässä käyttämiämme menetelmiä ja saamiamme ensimmäisiä tuloksia.

Mauissa käytetään muiden algoritmisten menetelmien ohessa ohjattua koneoppimista, jossa järjestelmä pyrkii rakentamaan tilastollisen mallin tekstiaineistosta ja tekstin kannalta olennaisista asiasanoista. Automaattisen annotoinnin testaamista varten olemassaoleva asiasanoitettu aineisto jaettiin opetus- ja testiaineistoon, joista ensinmainitun perusteella luotiin tilastollinen malli. Tämän jälkeen testiaineistoon kuuluvat asiakirjat asiasanoitettiin Mauin avulla käyttäen opetusaineistosta luotua mallia, ja näin saatua asiasanoitusta verrattiin normien alkuperäisiin asiasanoihin.

4.2 Arviointimenetelmät

Asiasanoituksen laadun arviointiin on useita eri mittausmenetelmiä. Yksinkertaisin toteuttaa on tiedonhaun piirissä käytetty tulosten tarkkuuden (precision) ja saannin (recall) mittaus. Tiedonhaussa tarkkuudella tarkoitetaan sitä, kuinka suuri osuus hakujärjestelmän tuottamista tuloksista on relevantteja, ja saannilla puolestaan sitä, kuinka suuren osuuden kaikista mahdollisista relevanteista tuloksista järjestelmä tuottaa. Tässä yhteydessä relevanteiksi tuloksiksi oletetaan alkuperäiset normitietokantaan tallennetut asiasanat, joten tarkkuus- ja saantiarvot mittaavat sitä, kuinka hyvin automaattisen järjestelmän tuottamat asiasanat keskimäärin täsmäsivät alkuperäisten ihmisen määrittämien asiasanojen kanssa. Automaattisen annotoinnin laatu ja sen arviointi riippuvat tällöin myös alkuperäisen metatiedon laadusta ja soveltuvuudesta oletettuun käyttötarkoitukseen. Tässä tutkimuksessa alkuperäisen asiasanoituksen laadun arvointi kuitenkin sivuutetaan.

Yksinkertaisen tarkkuuden ja saannin mittauksen ongelmana asiasanoituksen arvioinnissa on myös se, ettei tarkoituksenmukaisten asiasanojen valinta ole yksikäsitteistä. Myös toimialan asiantuntijoiden samalle tekstiaineistolle valitsemien asiasanojen välillä on suurta vaihtelua. Annotoijien keskimääräistä yhdenmukaisuutta voidaan mitata esimerkiksi Rollingin menetelmällä [10]. Esimerkiksi suomenkielisiä sosiaali- ja terveysalan aineistoja annotoitaessa asiantuntijoiden keskimääräiseksi yhdenmukaisuudeksi on aiemmin havaittu 33,7% [11].

Normitietokannan aineistoa koskevat tulokset on saatu kymmenkertaisella ristiinvalidoinnilla, jossa aineisto jaettiin ensin kymmeneen yhtäsuureen osaan. Näin saatiin kymmenen testikierrosta, joista kullakin testiaineistona käytettiin yhtä kymmenesosaa ja opetusaineistona vastaavasti muita yhdeksää osaa aineistosta. Lopullisiksi tarkkuus- ja saantiarvoiksi laskettiin kaikkien kymmenen annotointikierroksen tulosten keskiarvo.

Opetus- ja testiaineistoista jätettiin pois normit, joihin ei liittynyt julkista saatavilla olevaa tekstiaineistoa tai joiden metatiedoissa oli määritelty vähemmän kuin kaksi asiasanaa. Alustavia kokeiluja tehtiin myös siten, että asiasanojen vähimmäismääräksi oli asetettu yksi, kolme tai neljä, mutta asettamalla alaraja kahteen asiasanaan saatiin lupaavimmat tarkkuus- ja saantiarvot. Asiakirjat, joille on määritelty liian vähän asiasanoja, ovat vähemmän hyödyllisiä opetus- tai testiaineistona, mutta toisaalta asettamalla rima

liian korkealle rajataan pois suuri määrä asiakirjoja, jolloin aineisto jää määrällisesti pieneksi. Raja-arvon asettaminen on siten kompromissi käytettävän opetus- ja testiaineiston laadun ja määrän välillä.

Lisäksi asiasanat, joille on ontologiassa useita mahdollisia merkityksiä, samastettiin testausvaiheessa yhdeksi termiksi, koska Maui ei aina pysty erottelemaan toisistaan monikäsitteisten termien eri merkityksiä [11].

4.3 Alustavat tulokset ja havainnot

Ristiinvalidoinnilla keskimääräiseksi tarkkuudeksi saatiin 29,4% ja saanniksi 13,4%. Vertaamme seuraavassa tuloksia aiemmissa Maui-asiasanoittajan arvioinneissa saatuihin tuloksiin ja esitämme tuloksista alustavia tulkintoja. Vertailun osalta on huomattava, että aiemmat tulokset on saatu yksi pois -ristiinvalidoinnilla, joten hieman erilaisesta mittausmenetelmästä johtuen tulokset eivät välttämättä ole täysin vertailukelpoisia. Esitämme ne kuitenkin tässä normien annotointitulosten asettamiseksi yleisempään kontekstiin.

Aiemmissa Mauin arvioinneissa on saatu aineistosta ja käytettävästä lemmausmenetelmästä riippuen tarkkuusarvoja ainakin väliltä 25-40% [11], joten normien asiasanoituksen tarkkuus jäi hieman aiemmista sosiaali- ja terveysalan aineistoilla saaduista tuloksista mutta sijoittui kuitenkin aiempien tulosten laajaan haarukkaan. Saanti jäi sensijaan merkittävästi aiempien mittausten tuloksista, joten suurin osa alkuperäisten asiasanoittajien määrittämistä normien annotoinneista jäi automaattisen annotoinnin tuottamien asiasanojen ulkopuolelle.

Eräs mahdollinen tekijä heikossa saantituloksessa on normien olemassaolevan asiasanoituksen suhteellinen harvuus. Monet asiasanat esiintyvät käyttämämme aineiston metatiedoissa vain muutaman kerran, jolloin opetusdatasta rakennetun tilastollisen mallin ennustusvoima jää heikoksi. Toisaalta voi olla, että normien asiasanoituksessa on käytetty muita aineistoja useammin aihetta kuvaavia asiasanoja, jotka eivät esiinny tekstissä itsessään, jolloin asiasanojen tekstistä eristämiseen perustuva asiasanoitusalgoritmi ei välttämättä löydä samoja asiasanoja. Tuloksen taustatekijöiden tarkempi analyysi vaatisi lisätutkimusta.

Tulosten tulkinta puoliautomaattisen asiasanoituksen käyttökelpoisuuden kannalta riippuu osittain sekä metatiedon käyttötarkoituksista että automaattisen asiasanoituksen roolista osana tiedon laadintaprosessia. Jos annotaatiojärjestelmän käyttöliittymä suunnitellaan siten, että asiakirjan laatijan on helppo valita automatiikan ehdottamien asiasanojen joukosta sopivat, annotoinnin lievä epätarkkuus ei välttämättä haittaa merkittävästi. Vastaavasti, jos automatiikan odotetaan tuottavan vain osan asiasanoituksesta ja nopeuttavan näin asiakirjan laatijan työtä, matala saanti ei välttämättä ole suuri ongelma.

Odotamme lisäarvioinnin tuloksia ennen tarkempia johtopäätöksiä.

5 Jatkotutkimus

Tässä luvussa tarkastellaan tehdyn tutkimuksen esiintuomia kysymyksiä ja ongelmia sekä esitetään mahdollisia tulevia tutkimussuuntia.

5.1 Automaattisen asiasanoituksen tarkempi arviointi

Saatujen alustavien tulosten lisäksi aiomme toteuttaa lähitulevaisuudessa automaattisen asiasanoituksen tarkemman arvioinnin, jossa puolustusvoimien sisäiset asiantuntijat arvioivat automaattisesti tuotettujen asiasanoitusten onnistumista.

Asiasanoituksen monikäsitteisyyden lisäksi laadun mittaamista vaikeuttaa se, että valittujen asiasanojen hyödyllisyys riippuu metatiedon käyttötarkoituksesta. Esimerkiksi tiedonhaussa metatiedon hyödyllisyys riippuu merkittävästi siitä, kuinka hyvin asiasanat ja todelliset hakutermit vastaavat toisiaan. Lisäksi suhteellisen kapea-alaiset asiasanat, jotka eivät suoraan esiinny itse tekstissä mutta kuvaavat asiakirjaa riittävällä täsmällisyydellä, voivat olla puhtaan tekstihaun kannalta hyödyllisempiä kuin yleisluontoiset asiasanat. Toisaalta tiedon luokittelussa myös yleisluontoisemmista asiasanoista voi olla hyötyä. Asiasanojen merkitys voi poiketa puhtaasta tekstihausta myös moninäkymähaussa, jossa niin asiasanoja kuin muitakin metatietoja voidaan käyttää hakutulosten tarkempaan rajaamiseen.

Asiasanoituksen merkitystä tiedon löydettävyyden kannalta on näin ollen työlästä mitata täsmällisesti. Web-pohjaista aineistoa koskeneessa tutkimuksessa on aiemmin havaittu, ettei asiasanoituksesta ollut tiedonhaun kannalta merkittävää hyötyä [4]. Toisaalta kirjoittajat toteavat web-sivujen välisten linkkien ja sivujen osoitteiden olevan haun kannalta hyödyllisiä metatietoja, ja perinteisen asiakirjallisen tiedon haussa tällaisia tietoja ei välttämättä ole käytettävissä. Lisäksi heidän tutkimissaan tapauksissa tiedontarpeen täyttämiseen riitti yksi hakutulos, joka tarjosi haetun tiedon. Asiakirjallisen tiedon haussa voi olla tarpeen löytää kaikki käsiteltävän asian kannalta olennaiset asiakirjat.

Asiasanoituksen laatua ja merkitystä asiakirjallisen tiedon löydettävyyden näkökulmasta voitaisiin tutkia esimerkiksi tiedonhaun tutkimuksen menetelmin todellisten tiedontarpeiden ja hakutapausten pohjalta tai suhteellisen pitkäkestoisen seurantatutkimuksen avulla.

6 Yhteenveto

Tässä artikkelissa olemme kuvanneet prosessin asiakirjallisen tiedon saattamiseksi semanttisessa webissä ja yhdistetyssä tiedossa käytettävään RDF-muotoon. Tapaustutkimuksena käsittelimme normitietokannan metatietojen muuntamista, yhdistämistä ontologiaan sekä tiedon mahdollista käyttöä yksinkertaisessa moninäkymähaussa. Muunnosta varten kehitettiin työkalu, josta voi olla hyötyä myös muiden SÄHKE2-metatietomallin mukaisten aineistojen muuntamisessa ja semanttisessa rikastamisessa.

Selvitimme myös vapaiden asiasanojen ja ontologisten käsitteiden välistä suhdetta ja normien puoliautomaattista asiasanoitusta. Testasimme viimeksimainittua käyttäen ARPA-rajapintaa ja Maui-annotoijaa asiasanoitusten tuottamiseen. Alustavissa tuloksissa havaittiin normien automaattisen asiasanoituksen tarkkuuden olevan samassa luokassa muilla aineistolla saatujen tulosten kanssa saannin jäädessä kuitenkin heikommaksi aiempiin tuloksiin nähden. Odotamme lisäksi lähitulevaisuudessa tehtävän lisäarvioinnin tuloksia. Tulosten merkitys ja puoliautomaattisen asiasanoituksen käyttökelpoisuus riippuvat asiasanoituksen käyttötarkoituksista ja metatietoa hyödyntävien järjestelmien toteutuksesta.

Semanttisen webin teknologioihin siirtymiseen tarvittavat työkalut ja menetelmät ovat suurelta osin olemassa. Asiasanastoista ja perinteisestä metatiedosta on mahdollista siirtyä ontologioiden ja yhdistetyn tiedon piiriin suhteellisen vaivattomasti. Tämän jälkeen semanttisen webin tekniikat mahdollistavat uudentyyppiset lähestymistavat olemassaolevan aineiston käsittelyyn ja hallintaan. Tiedon esittäminen toisiinsa viittaavina resursseina ja niiden muodostamana verkostona mahdollistaa koneellisen loogisen päättelyn ja siihen perustuvat sovellukset kuten semanttiset haku- ja suosittelujärjestelmät. Looginen seuraava askel onkin kartoittaa näiden sovellusten mahdollisuuksia ja hyötyjä käytännön asiakirjahallinnossa.

Kiitokset Tämä työ on tehty osana FinnONTO – Suomalaiset semanttisen webin ontologiat (2003-2012) -hanketta, jota tällä hetkellä rahoittavat Teknologian ja innovaatioiden kehittämiskeskus Tekes ja 35 julkisen alan organisaatiosta ja yrityksestä koostuva konsortio.

Viitteet

1. Tim Berners-Lee, James Hendler, and Ora Lassila. The Semantic Web (Berners-Lee et. al 2001). May 2001.
2. Matias Frosterus, Eero Hyvönen, and Mika Wahlroos. Extending ontologies with free keywords in a collaborative annotation environment. In *Proceedings of the ISWC 2011 Workshop Ontologies Come of Age in the Semantic Web (OCAS)*. CEUR Workshop Proceedings, Vol 809, http://ceur-ws.org, ISSN 1613-0073, October 2011.
3. Thomas R. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199 – 220, 1993.
4. David Hawking and Justin Zobel. Does topic metadata help with web search? *Journal of the American Society for Information Science and Technology*, 58(5):613–628, 2007.
5. E. Hyvönen, S. Saarela, and K. Viljanen. Application of ontology-based techniques to view-based semantic search and browsing. In *Proceedings of the First European Semantic Web Symposium, May 10–12, Heraklion, Greece*. Springer–Verlag, 2004.
6. Eero Hyvönen. Miksi asiasanastot eivät riitä vaan tarvitaan ontologioita?, Oct 2005.
7. Jussi Kurki and Eero Hyvönen. Collaborative metadata editor integrated with ontology services and faceted portals. In *Workshop on Ontology Repositories and Editors for the Semantic Web (ORES 2010), the Extended Semantic Web Conference ESWC 2010, Heraklion, Greece*. CEUR Workshop Proceedings, http://CEUR-WS.org, 2010.
8. Joonas Laitio. Semantic web data quality control. Master's thesis, Aalto University, School of Electrical Engineering, Degree Programme of Automation and Systems Technology, October 2011.
9. Olena Medelyan. *Human-competitive automatic topic indexing*. PhD thesis, University of Waikato, July 2009.
10. Loll N. Rolling. Indexing consistency, quality and efficiency. *Information Processing & Management*, 17(2):69–76, 1981.
11. Reetta Sinkkilä, Osma Suominen, and Eero Hyvönen. Automatic semantic subject indexing of web documents in highly inflected languages. In *Proceedings of the 8th Extended Semantic Web Conference (ESWC 2011)*, June 2011.
12. Pasi Tapanainen and Timo Järvinen. A non-projective dependency parser. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, 1997.