KATSAUS

Automaattinen asiasanoitus Radio- ja televisio-ohjelmatietokanta Ritvassa

Tommi Lehtonen
Tekninen suunnittelija
Kansallinen audiovisuaalinen instituutti
tommi.lehtonen@kavi.fi

Juha Piukkula
Erikoissuunnittelija
Kansallinen audiovisuaalinen instituutti
juha.piukkula@kavi.fi

The National Audiovisual Institute's (KAVI) radio and television archive started a joint project with the Finnish broadcasting company (Yle) and the National Library of Finland to develop automated indexing using program subtitles as a data source. The project relies on Annif, a tool originally developed by Osma Suominen. Annif is built upon a combination of existing natural language processing and machine learning tools. It is designed to be multilingual and it can support any subject vocabulary. Annif can use several different backends. During the spring and summer of 2019, 313 Yle programmes were jointly annotated by KAVI and Yle for Annif testing. Analysis was performed through cross-validation of manual and automated results. It was noted that a television programme may be produced so that the central theme is not verbally mentioned at all, making purely subtitle-based classification hard. When a brief programme description was included, the results improved. The results and quality of Annif-based indexing were promising and the project will continue.

Keywords: automaattinen sisällönkuvailu; asiasanoitus; sisällönkuvailu; koneoppiminen; ohjelmatekstitys; muistiorganisaatiot; audiovisuaalinen aineisto

Pysyvä osoite: https://doi.org/10.23978/inf.88107

Johdanto

Digitalisaation mahdollisuuksia ja haasteita on pohdittu usean vuoden ajan muistiorganisaatioissa (ks. esim. Harju & al. 2018, 9). Yksi tutkituimmista aihepiireistä on kulttuuriaineiston sisällönkuvailun automatisointi, johon yksi syy on aineistomäärien nopea kasvu muistiorganisaatioissa. Henkilökunnan määrä on rajallinen, eikä ehdi kuvailemaan aineistoa siinä määrin kuin olisi tarpeen. Tässä artikkelissa tutkitaan automaattisen asiasanoituksen käyttöönottoa Kansallisen audiovisuaalisen instituutin (KAVI) tuottamassa radio- ja tv-arkistotietokannassa eli Ritvassa. Sen aineistomäärä kasvaa lähes neljällä tuhannella ohjelmalla päivässä.

KAVI on opetus- ja kulttuuriministeriön alainen valtion virasto. KAVIn lakisääteisiin tehtäviin kuuluvat elokuvien ja televisio- ja radio-ohjelmien säilyttäminen sekä niihin liittyvä tutkimus, kuvaohjelmien tarjoamisen valvonta ja mediakasvatuksen edistäminen. Kansallinen audiovisuaalinen arkisto (KAVA vuosina 2008–2013, Suomen elokuva-arkisto vuosina 1957–2007) ja Mediakasvatus- ja kuvaohjelmakeskus (MEKU vuosina 2012–2013, Valtion elokuvatarkastamo vuosina 1946–2011) yhdistyivät 1.1.2014 Kansalliseksi audiovisuaaliseksi instituutiksi.

Radio- ja televisioarkisto (RTVA) on perustettu huolehtimaan suomalaisen radio- ja tv-kulttuurin säilymisestä jälkipolville. Katselu- ja kuuntelupisteissä voi tutkia vuoden 2009 alusta lähtien tallennettua ohjelmavirtaa sekä Yleisradion tuottamia tv-ohjelmia ajalta 1957–2008. Vuoden 2008 alussa perustetun radio- ja tv-arkiston toiminta perustuu kulttuuriaineistolakiin, joka asettaa velvoitteita sekä radio- ja tv-arkistolle että ohjelmien tuottajille. RTVA palvelee tutkijoita ja opiskelijoita tallentamalla televisio- ja radiokanavien digitaalista ohjelmavirtaa, jota voi selata ja tutkia katselu- ja kuuntelupisteissä seitsemällä eri paikkakunnalla. Lisäksi RTVA tallettaa laissa määritellyt ohjelmatyypit alkuperäistasoisina fyysisinä kappaleina tai tiedostoina. (Kansallinen audiovisuaalinen instituutti, 2019a.)

Ritva sisältää keskeisten kotimaisten radio- ja tv-kanavien ohjelmatiedot ja ohjelmavirran kokonaisuudessaan sekä näytteitä muilta kanavilta. Ohjelmien tallennus on aloitettu vuoden 2009 alussa. Tallennettaessa ohjelmatietoja Ritvaan käytetään hyväksi eri tietolähteitä. Tärkeimpiä näistä ovat Venetsia-ohjelmatietojärjestelmä, Finnpanel, EPG, Yle ja Radiot.fi. Eri kanavilta saadaan ohjelmatietoja myös suoraan. Ritva-tietokannan metatieto on vapaasti internetissä selailtavana. Tietokannan suoratoistoihin on pääsy KAVIssa, Kansalliskirjastossa, Eduskunnan kirjastossa, Turun, Jyväskylän, Oulun ja Itä-Suomen yliopistojen kirjastoissa, Åbo Akademin kirjastossa sekä Tampereen yliopistossa.

Keväällä 2019 aloitettiin projekti, jonka tarkoituksena on kehittää Ritvan automaattista asiasanoitusta käyttäen lähteenä ohjelmatekstityksiä (suomenkieliset ja kuulovammaisille tarkoitetut). Sovelluksena käytetään alkujaan Osma Suomisen kehittämää Annif-työkalua, jonka jatkokehitykseen on Kansalliskirjastossa sitouduttu (Suominen, 2019c). Tarkoituksena on hyödyntää asiasanoja ohjelma-aineiston sisällönkuvailussa, jotta tutkijat ja opiskelijat löytäisivät haluamaansa tutkimusaineistoa entistä paremmin. Projektissa ovat mukana KAVI ja Yle Arkisto. Tämä artikkeli esittelee alustavia tuloksia ohjelmiston käytöstä ja suosituksia jatkotoimenpiteistä.

Menetelmät

Annif tuottaa automaattista asiasanoitusta tekstianalyysin perusteella, tarkoituksena helpottaa ihmisvoimin työlästä asiasanoittamista käyttäen olemassa olevaa metadataa (Suominen, 2019a, 1–2). Ohjelmisto on avointa koodia. Annifin on todettu soveltuvan parhaiten asiatekstille, mutta työkalua testataan myös muilla aineistoilla. Ideana on antaa ohjelmistolle teksti, jonka käsiteltyään Annif ehdottaa asiasanoja tekstille.

Ohjelmisto pohjautuu yhdistelmään olemassa olevia luonnollisten kielten käsittely- ja koneoppimistyökaluja. Työkalu on suunniteltu monikieliseksi. Annifissa voi käyttää useita eri taustaohjelmia, toisin sanoen erilaisia algoritmeja. Niitä pitää opettaa dokumenteilla, jotka on asiasanoitettu johonkin sanastoon pohjautuen. Taustaohjelmia voi lisätä Annifin käyttöön vapaasti. Pohja-aineistona on mm. Finnasta poimittua, asiasanoitettua materiaalia.

Annifin arkkitehtuuri lyhyesti

Annif on toteutettu Python-ohjelmointikielellä. Aiheen indeksointi ja luokittelu käsitellään taustaohjelmissa, joita käytetään joko yksinään tai yhdessä. Asetukset määritellään tiedostoon (projects.cfg) ja kouluttaminen tehdään komentorivillä (Suominen, 2019c).

Annifin käyttö aloitetaan määrittelemällä projektitiedosto. Asetuksiin laitetaan käytettävä sanasto ja taustaohjelmakohtaiset parametrit. Koulutusaineisto on esikäsiteltävä ennen kuin sitä voidaan analysoida. Annifissa tekstin esikäsittelyä hoitavat Analyzer-moduulit, jotka muuttavat tekstin ohjelmalle ymmärrettäviksi lauseiksi ja yksittäisiksi sanoiksi. Sanat voidaan edelleen normalisoida käyttämällä kielikohtaisia algoritmeja perusmuotoistamiseen (Wahlroos, 2013, 9).

Annifin sanastomoduuli käsittelee sanastotietojen lataamista ja tallentamista. Tallentamiseen käytetään joko TSV- tai SKOS / RDF-tiedostoja. Sama sanasto voidaan jakaa useiden Annif-projektien kesken.

Annifin käyttämät taustaohjelmat

Annifin ideana on koota yhteen erilaisia koneoppimisen metodeja ja taustaohjelmia ja käyttää niitä yhdessä parhaan mahdollisen lopputuloksen saamiseksi. Suominen jakaa automaattisen asiasanoituksen tekniikat kahteen ryhmään, leksikaalisiin ja assosiatiivisiin (ks. Pouliquen & al. 2003 ja Toepfer & al. 2018). Leksikaaliset menetelmät toimivat pääosin vertaamalla termejä sanastosta ja dokumenteista. Assosiatiiviset menetelmät käyttävät sanojen tai niiden osien korrelaatioita ja pohjautuvat suureen koulutusdataan. Koneoppiminen on keskeisempää assosiatiivisissa menetelmissä, mutta sitä hyödynnetään myös leksikaalisissa. Seuraavaksi esitellään lyhyesti tässä projektissa käytettyjä taustaohjelmia.

TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) on piirteenirrotusmenetelmä, joka laskee sanoille painokertoimia ja määrittää näin olennaiset piirteet kullekin luokalle.

TF-IDF hakee korrelaatioita käytettävän sanaston ja käytettävien dokumenttien välillä. Koulutusaineistona käytettävät tekstit on asiasanoitettu manuaalisesti. Näiden avulla lasketaan TF-IDF-arvot, jotka tallennetaan jokaiselle sanalle indeksiin. TF-IDF-menetelmässä käsiteltävää tekstiä arvioidaan laskemalla termeille painokertoimia, mikä tapahtuu laskemalla termien esiintymistiheyden summa ja termien esiintymisestä kaikissa dokumenttinäytteissä kertova luku. Saadun painokertoimen arvo on sitä suurempi, mitä harvinaisempi sana on. (Koskimies, 2017, 5–6, ks. myös Leuhu, 2014.) Taustaohjelmaan ei ole mahdollista vaikuttaa muutoin kuin koulutusaineistoa parantamalla.

fastText

fastText on Facebookin kehittämä ilmainen avoimen lähdekoodin kirjasto, jonka avulla voidaan luokitella tekstiaineistoja. Se käyttää sekä ohjattua että

ohjaamatonta oppimista. fastText on neuroverkon kaltainen malli. (Ks. esim. Joulin & al. 2017.)

Annif käyttää fastTextin ohjatun oppimisen malleja, koska Annif koulutetaan aina tietyllä aineistolla ja sanastolla. Koulutusaineistojen käsittelyssä voidaan antaa erilaisia parametreja. Näitä kertoja lisäämällä saadaan usein parempia tuloksia. Oppimistahtia muuttamalla voidaan vaikuttaa siihen, kuinka paljon muutoksia tehdään edelliseen aineiston läpikäyntiin verrattuna. Opetusaineisto jaetaan peräkkäisiin merkkijonoihin, joita kutsutaan N-grammeiksi. Näiden minimi- ja maksimimerkkimääriä voidaan muutella. (Facebook Research, 2019.)

Maui

Maui on leksikaalinen automaattisen asiasanoituksen algoritmi, joka on kehitetty Waikaton yliopistossa. Maui valitsee ehdotukset tekstin aiheiksi ja arvottaa aiheet niiden ominaisuuksien mukaan. Ominaisuudet voivat olla tilastollisia, semanttisia tai tietosanakirjamaisia. Taustaohjelma yhdistelee paljon erilaisia heuristisia ratkaisuja sanaston termien ja dokumentin aiheiden yhdistämiseksi ja tunnistamiseksi. (Medelyan, 2009, 108–109.)

Aiheet voivat olla asiasanoja tai artikkelien nimiä. Maui toimii hyvin pitkien tekstien kanssa ja löytää käytettävän sanaston mukaan tekstin aiheita. Mauin koulutukseen riittää muutama sata dokumenttia (ks. esim. Medelyan, 2009, 169 ja Suominen, 2019a, 6).

PAV

Annifin erikoisuutena on eri taustaohjelmien älykäs yhdistely. Annifissa voi käyttää yksinkertaista, eri taustaohjelmien antamien aihekohtaisten pisteiden keskiarvoihin perustuvaa yhdistelmää. Tässä projektissa päädyttiin kokeilemaan älykkäämpää muita taustaohjelmia yhdistävää taustaohjelmaa, jota pitää kouluttaa myös itsessään. (Suominen, 2019a, 9.)

Eri taustaohjelmien asiasanoille antamat tulokset painotetaan uudestaan käyttämällä isotonista regressiota. Siinä lopputulos esitetään laskemalla pistekeskiarvot jokaiselle pisteelle erikseen. Tavallisessa regressiokaaviossa keskiarvo esitetään suorana viivana. Annif käyttää tässä PAV-algoritmia (Pool adjacent violators) (Pedregosa, 2013; Suominen, 2019a, 9).

nn_ensemble

nn_ensemble on toinen tapa yhdistää taustaohjelmien tuloksia. nn_ensemble on neuroverkko, joka on tehty käyttäen TensorFlow-ohjelmointialustaa koneoppimisohjelmistoille ja Keras-nimistä Python-ohjelmakirjastoa. nn_ensemble-projektin parametreja säädetään. Esimerkiksi neuroverkon piilokerroksen neuronien määrää eli neuroverkon kokoa (nodes) voidaan muuttaa. Hävikkiarvoa (dropout_rate) säädellään, jotta lopputulos pysyisi riittävän yleispätevänä. Opetuskertojen läpikäyntien määrää (epochs) voidaan muuttaa kuten fastTextissä. nn_ensembleen on toteutettu mahdollisuus opettaa taustaohjelmaa edelleen (Suominen, 2019b).

Aineisto

Kokeilun aineistona käytettiin asiasanoitettuja Ylen asiaohjelmia ja niiden tekstityksiä. Asiasanoituksella viitataan ohjattuun asiasanoitukseen. Sillä tarkoitetaan aineiston kuvailua siten, että käyttäjät liittävät aineistoon tunnisteita ylläpidetystä tunnisteluettelosta (Sanastokeskus ry., 2011). Kansalliskirjaston ylläpitämä Yleinen suomalainen ontologia (YSO) on tällainen tunnisteluettelo. ”YSO on kolmikielinen, etupäässä yleiskäsitteistä koostuva ontologia. YSO on rakennettu suomalaisen kulttuuripiirin sisällönkuvailutarpeiden ja käsitteistön pohjalta, ja se on tarkoitettu käytettäväksi kuvailuun erityisesti silloin, kun kuvailtavien aineistojen aihealueet ovat monipuolisia” (Finto, 2020). Ontologia kuvaa aihealueeseensa kuuluvat käsitteet ja niiden väliset suhteet formaalisti koneluettavassa muodossa. Ontologiaa voidaan käyttää koneellisessa päättelyssä, toisin kuin asiasanastoa. (Asiasanastot ja ontologiat, 2020.)

Annifin käyttämät taustaohjelmat tarvitsevat malliksi valmiiksi tietyllä sanastolla asiasanoitettuja aineistoja. KAVIssa käytetään YSO-ontologiaa, joka on sama kuin Kansalliskirjastossa ja Finnassa on käytössä. YSO ei ole ainoa standardoitu ontologia joka toimii Annifilla. Yle käyttää useita eri ontologioita, joista KOKO-ontologia on yksi. KOKOn asiasanoilla on suurimmaksi osaksi YSO-ontologiassa vastineet (Finto, 2020b).

Sähköisen viestinnän palveluista annetussa laissa (Laki sähköisen viestinnän palveluista, 2014) on säädetty ääni- ja tekstitysvelvoitteesta televisio-ohjelmissa. Laissa mainitaan, että yleisen edun kanavilla on velvollisuus tarjota ääni- ja tekstityspalveluita myös kotimaisiin ohjelmiin. Laki velvoittaa Yleisradiota, MTV Mediaa, Nelosta ja AlfaTV:tä (Traficom, 2019). Tekstityk-

siä analysoimalla on mahdollista tuottaa automaattisesti asiasanaehdotelmia näiden yhtiöiden kanavien ohjelmiin.

KAVI saa kokoelmiinsa Yleltä heidän omatuotantoisiin ohjelmiinsa liittyvät, kuulovammaisille tarkoitetut tekstitykset. Projektin alussa asiasanoitettuja ohjelmia ei kuitenkaan löytynyt Yleltä eikä Ritvasta tarpeeksi. Kevään ja kesän 2019 aikana asiasanoitettiin RTVAn ja Yle Arkiston henkilökunnan sekä ohjelmantekijöiden yhteisvoimin 313 Ylen asiaohjelmaa vuosilta 2009–2018 Annifin testausta varten. RTVAn henkilökunta on saanut kirjastoalan koulutuksen ja Ylen puolesta asiasanoitukseen osallistuivat ohjelmien tekijät. Ylellä asiasanoitus tehtiin Areenan taustajärjestelmä Feenixiin, josta asiasanat siirrettiin rajapintojen kautta KAVIin. Ylen antamista asiasanoista poimittiin ne, joiden alkuperä on Finto-palvelun KOKO-ontologia. Asiasanat konvertoitiin YSO-muotoon, mikäli KOKOn asiasanoille löytyi vastine YSOsta.

Asiasanat siirrettiin Annifin koulutusta varten TSV- ja KEY-tiedostoihin. Näiden piti olla UTF-8-merkistökoodattuja ja kaikki turhat tekstit kuten aikakoodit ja tekstitysmuotoilut poistettiin tekstityksistä. Kuvan 1 prosessikaaviossa on hahmoteltu yhteistyön eri vaiheet.

Kuva 1. KAVIn ja Ylen ohjelmien asiasanoitusprojekti kaaviona.

Osma Suomisen ohjeen mukaan (Suominen, 2019d) testauksessa käytettiin ristiinvalidointitekniikkaa (cross-validation). Alkuperäinen näyte jaetaan satunnaisesti viiteen samankokoiseen alinäytteeseen. Viidestä ”aliosanäytteistä” yksi osanäyte säilytetään validointitietoina mallin testaamiseksi, ja loput käytetään harjoitustietoina. Prosessi toistetaan viisi kertaa, kutakin alinäytteistä käytetään vain kerran validointitietoina. Tuloksista lasketaan keskiarvo. Menetelmän etuna on, että samaa aineistoa voidaan käyttää sekä koulutukseen että validointiin. Tulokseksi saadaan kaikkien suorituskertojen keskiarvo (Wahlroos, 2013, 24).

Esimerkiksi alinäytteessä yksi käytettiin Maui-taustaohjelman koulutukseen 249 ohjelman asiasanoja, ja validointiin loppuja 64 ohjelmaa. Alinäytteessä kaksi taustaohjelma koulutettiin 250 ohjelmalla, jotka sisälsivät edelliseen validointiin käytetyt ohjelmat. Validointiin käytettiin 63 ohjelmaa, jotka olivat alinäyte yhden koulutusaineistossa. Periaatteena oli, ettei eri alinäytteiden koulutusaineistoissa saanut olla yhtään samaa ohjelmaa.

Jokaisesta taustaohjelmayhdistelmästä kerättiin arvot numeerisina väliltä 0 ja 1, jossa 1 on paras ja 0 huonoin.

P@5 eli tarkkuus (Precision): laskettiin desimaaliluku sille, kuinka hyvin samat viisi sanaa löytyivät testiaineistosta joita taustaohjelma ehdotti. Esimerkiksi oletetaan, että on kolme ohjelmaa, joille pitäisi kullekin löytää viisi asiasanaa eli yhteensä 15 asiasanaa. Taustaohjelman ehdotusten kärjessä on yhteensä kuusi samaa asiasanaa. Tarkkuus on tällöin 6/15 eli 0,40. (Wahlroos, 2013, 19–20.) Tarkkuus lasketaan viiden (@5) eniten pisteitä saaneen sanan perusteella.

F1 score doc average@5: F1 on tarkkuuden (Precision, ks. edellä) ja saannin (Recall) painotettu harmoninen keskiarvo. Saanti on palautettujen relevanttien asiasanojen osuus kaikista kyselyn kannalta olennaisista asiasanoista. Oletetaan, että edellä esitetyssä esimerkissä kaikkien relevanttien sanojen summa on 20. Näistä taustaohjelma ehdottaa 15:ta sanaa, joista kuusi on samoja. Saanti on tällöin 6/20 eli 0,3. (Wahlroos, 2013, 19–20). Tarkkuuden ja saannin harmonisen keskiarvon kaava on (kun p = tarkkuus ja r = saanti) F1 = (2*pr)/p+r. (Wahlroos, 2013, 25). F1 lasketaan viiden (@5) eniten pisteitä saaneen asiasanan perusteella.

NDCG@5 (Normalized Discounted Cumulative Gain): tässä tekniikassa dokumentin sijalukua tuloslistassa käytetään hyödyn arvoa alentavana tekijänä. Mitä suurempi on dokumentin sijaluku listalla, sitä pienempi on dokumentin kumuloituun hyötyyn lisäämä arvo. Listan kärkipäässä olevat sanat saavat siis lisäpisteitä. Nämä pisteet yhdistetään dokumentin relevanssipisteisiin. (Järvelin & al. 2000.)

Annifin taustaohjelmat antavat relevanssipisteet. NDCG-luku lasketaan viiden (@5) eniten pisteitä saaneen sanan perusteella.

Taulukossa 1 näkyvät testauksessa käytetyt taustaohjelmat ja niiden yhdistelmät. TDF-IDF ja fastText on koulutettu Finna-aineiston versio Cicerolla. fastTextin parametrit on optimoitu Cicero-versioon. Maui, PAV ja nn_ensemble on koulutettu tekstitysaineistolla.

Taustaohjelmat	Yhdistelmät
TF-IDF	-
AfastText	-
Maui	-
TF-IDF+fastText+Maui	PAV
TF-IDF+fastText+Maui	nn_ensemble

Taulukko 1. Käytetyt taustaohjelmat ja yhdistelmät

Tekstitysten lisäksi Ritvaan tulevat myös ohjelmien kuvaukset. Kokeilimme, mitä tapahtuu, kun yhdistetään kuvaukset tekstitysaineistoon ja yllä kuvailtu operaatio uusitaan. Aineistojen eval-tuloksia verrattiin eri taustaohjelmien kesken. Lisäksi tehtiin sanoitustestaus. PAV- ja nn_ensemble-yhdistelmissä käytettävien taustaohjelmien keskinäinen painotus on yhtä suuri.

Tulokset

Testiaineistoja testattiin Annifin eval-komennolla. Tuloksia käytiin läpi taustaohjelma kerrallaan. Taulukkoon 2 on laskettu alinäytteiden keskiarvot. Taustaohjelmien antamien asiasanojen laatua tarkasteltiin ottamalla aineiston eri alinäytteistä sattumanvaraisesti neljä esimerkkiohjelmaa.

Validointi

Taustaohjelmille tehtiin validointi aikaisemmin esiteltyjen mittareiden mukaisesti. Alinäytteet analysoitiin erikseen ja niiden tuloksista laskettiin aritmeettinen keskiarvo.

Taulukko 2 näyttää, kuinka kaikilla taustaohjelmilla kuvausten lisäys aineistoon tuotti selvästi parannusta tuloksiin. Maui toimii aineistossa hyvin, paremmin kuin kumpikaan tilastollisista taustaohjelmista. On kuitenkin huomioitava, että ne oli koulutettu Finna-aineistolla ja Maui pelkästään tekstitysaineistolla.

Tausta-ohjelma	**Testiaineisto**	**P@5** Tarkkuus, jos on viisi samaa asiasanaa	**F1@5** (jatkossa F1 viittauksissa) Tarkkuuden ja saannin keski-arvo	**NDCG@5** Painotettu tulos, ensim-mäisestä saa eniten pisteitä, toisesta toisek-si eniten jne.
tfidf	Tekstitykset	0,1303	0,1015	0,1583
tfidf	tekstitykset + kuvaukset	0,1375	0,1164	0,1770
fasttext	Tekstitykset	0,1818	0,1410	0,2060
fasttext	tekstitykset + kuvaukset	0,1903	0,1452	0,2147
maui	Tekstitykset	0,2923	0,2243	0,3404
maui	tekstitykset + kuvaukset	0,3322	0,2555	0,3901
PAV	Tekstitykset	0,3027	0,2335	0,3535
PAV	tekstitykset + kuvaukset	0,3277	0,2493	0,3844
nn_ensemble	Tekstitykset	0,3747	0,2910	0,4236
nn_ensemble	tekstitykset + kuvaukset	0,4071	0,3174	0,4664

Taulukko 2. Analyysin tulokset

fastText	PAV	nn_ensemble
analyzer=snowball(finnish) dim=360 lr=0.5 epoch=140 minCount=5 minn=5 maxn=8 loss=hs limit=1000 chunksize=24 vocab=yso-fi	min-docs=1 (ottaa jokaisen asiasanan huomioon) limit=100 vocab=yso-fi	limit=100 vocab=yso-fi nodes=100 dropout_rate=0.2 epochs=30

Taulukko 3. Projekteissa käytetyt parametrit

Sanoitustestaus

Miten automaattinen asiasanoitus sitten vertautuu RTVAn sisällönkuvailukoulutuksen saaneen henkilökunnan tekemään asiasanoitukseen? Yksi koneoppimisen keskeisistä haasteista on se, että taustaohjelmat koulutetaan kovin poikkeavilla piirteillä. Asetukset tulisi saada sellaisiksi, että ne ovat mahdollisimman hyvin yleistettävissä kaikkeen aineistoon. Myös asiasanoituksen laatua pitää tarkastella. Tässä laadulla tarkoitetaan automaattisen sanoituksen asiasanoituksen osuvuutta sisältöön nähden sekä vastaavuutta ihmisten antamiin asiasanoihin.

Testauksessa oli neljä esimerkkiohjelmaa eri alinäytteistä. Testissä käytettiin PAV- ja nn_ensemble -yhdistelmiä. Annettuja asiasanoja verrattiin RTVAn henkilökunnan antamiin asiasanoihin. Ehdotettuja asiasanojen määrä rajattiin viiteen, kuten oli tehty validointitesteissäkin. Kaikki ohjelmat löytyvät Ritva-tietokannasta (Kansallinen audiovisuaalinen instituutti, 2019b).

Ohjelmassa ”Syötävät sävelet: Kissaherra Ravel” puhuttiin varsin vapaamuotoisesti säveltäjä Maurice Ravelin ruokatottumuksista kahden henkilön järjestämällä piknikillä. Esiintyjien keskustelu oli polveilevaa. Taulukossa 4 näkyy, kuinka PAV-taustaohjelma antaa pelkillä tekstityksillä koulutettuna kaksi samaa asiasanaa kuin asiasanoittaja oli antanut. ”Ensimmäinen maailmansota” on ohjelman kannalta relevantti. Asiasanat ”sarjakuvataiteilijat” ja ”kodinkoneet” sen sijaan eivät liity ohjelmaan millään tavoin. Kun koulutusaineistoon ja analysoitavaan aineistoon lisättiin ohjelman kuvaus, asiasanan ”kodinkoneet” tilalle tuli paljon relevantimpi ”lihansyönti”. Huomattava parannus nähtiin käytettäessä nn_ensemblea: kaikki ehdotetut asiasanat olivat myös asiasanoittajan antamia.

Luetteloijan antamat	**PAV (tekstitykset)**	**PAV (tekstitykset ja kuvaus)**	**nn_ensemble (tekstitykset)**	**nn_ensemble (tekstitykset ja kuvaus)**
kulttuurihistoria kulinarismi säveltäjät ruokalajit syöminen juomat ruoanvalmistus	säveltäjät ruoanvalmistus sarjakuvataiteilijat ensimmäinen maailmansota kodinkoneet	säveltäjät ruoanvalmistus ensimmäinen maailmansota sarjakuvataiteilijat lihansyönti	Säveltäjät ruoanvalmistus kulttuurihistoria juomat syöminen	säveltäjät ruoanvalmistus syöminen kulttuurihistoria juomat

Taulukko 4. Syötävät sävelet: Kissaherra Ravel -ohjelman asiasanoitus
Alinäyte: 5, ID: PROG_2017_00717037
Yle Teema & Fem , 15.8.2017 klo 20.50-20.59
https://www.rtva.kavi.fi/program/details/program/25683811

”Musiikin voima : Mauno Järvelä” -ohjelma on kohteena olevan viulistin ja pelimannimuusikon monologi taiteesta ja musiikista. Taulukossa 5 nähtäviin taustaohjelmien antamiin kehnohkoihin tuloksiin selityksenä on se, että ilmaisu on hyvin visuaalista. Asiasanoittaja on käyttänyt ennakkotietoaan henkilöstä. Ohjelmassa ei mainita viuluista tai pelimannimusiikista mitään, mutta kyseinen henkilö on viulisti ja tunnettu pelimannimuusikko. PAV ei anna yhtään samaa asiasanaa kuin asiasanoittaja.

Asiasana ”apuvälineet” ei liity sisältöön mitenkään, eikä ohjelmassa käsitellä Musiikin aika -tapahtumaa. nn_ensemble ei myöskään anna yhtään samaa asiasanaa kuin asiasanoittaja. ”Ruoanvalmistus” sekä ”ruoat (ruokalajit)” eivät ole ohjelmassa relevantteja. Muut taustaohjelman antamat asiasanat ovat osittain relevantteja.

Luetteloijan antamat	**PAV (tekstitykset)**	**PAV (tekstitykset ja kuvaus)**	**nn_ensemble (tekstitykset)**	**nn_ensemble (tekstitykset ja kuvaus)**
lapset (ikäryhmät) kansansoittajat musiikinopettajat pelimannimusiikki viulumusiikki viulistit	säveltäjät apuvälineet nykymusiikki äänimaisema Musiikin aika	taidemusiikki viestintä taiteet säveltäjät teologit	säveltäjät äänimaisema ruoanvalmistus hiljaisuus ruoat (ruokalajit)	viestintä taide säveltäjät taiteet ruoanvalmistus

Taulukko 5. Musiikin voima : Mauno Järvelä -ohjelman asiasanoitus
Alinäyte: 4, ID: PROG_2017_00728430
Yle Teema & Fem , 10.12.2017 klo 16.45-16.51
https://www.rtva.kavi.fi/program/details/program/26791648

”Suomalaisia sarjakuvataiteilijoita” -ohjelmassa sarjakuvapiirtäjät kertovat omista töistään, mutta ohjelman aihetta eli sarjakuvia ei välttämättä mainita ohjelmissa kertaakaan. Taulukosta 6 näkyy, kuinka kuvaukset lisäämällä asiasanoitus parani huomattavasti. PAV tuotti epärelevantin asiasanan ”musiikkivideot” kummassakin aineistoversiossa. Kun ohjelman kuvaus ei ollut mukana, nn_ensemble antoi täysin epärelevantin asiasanan ”Puerto Rico”. Kuvaus lisättynä nn_ensemble antoi kolme samaa asiasanaa kuin asiasanoittaja. Loput kaksi asiasanaa olivat relevantteja.

Luetteloijan antamat	**PAV (tekstitykset)**	**PAV (tekstitykset ja kuvaus)**	**nn_ensemble (tekstitykset)**	**nn_ensemble (tekstitykset ja kuvaus)**
sarjakuvat sarjakuvataiteilijat kerronta juoni tarinat kertomukset ideat muisti (kognitio) groteski	piirustusvälineet tukeminen tarinat musiikkivideot kertoja	taidemusiikki viestintä taiteet säveltäjät teologit	säveltäjät äänimaisema ruoanvalmistus hiljaisuus ruoat (ruokalajit)	viestintä taide säveltäjät taiteet ruoanvalmistus

Taulukko 6. Suomalaisia sarjakuvataiteilijoita -ohjelman asiasanoitus
Alinäyte: 4, ID: PROG_2018_00745968
Yle Teema & Fem , 11.7.2018 klo 16.51-16.53
https://www.rtva.kavi.fi/program/details/program/28949057

”Mun heimo” -ohjelman kesto on puoli tuntia. Asiasanoittaja on antanut paljon asiasanoja, koska ohjelmassa on monta aihetta. PAV tuotti asiasanan ”hyllyt”, koska henkilöt hyllyttivät tavaroita kaupassa. ”Hyllyt”-asiasanaa ei kuitenkaan voi pitää ohjelman keskeisten teemojen kannalta relevanttina. Saman voi todeta nn_ensemblen antamasta asiasanasta ”porot”. Taulukosta 7 näkee, että vain ”somalit” on sama molemmilla taustaohjelmilla, kun käytössä olivat kuvaus ja tekstitykset. nn_ensemble tuotti täsmälleen samat relevantit asiasanat, oli kuvaus mukana tai ei.

Luetteloijan antamat	**PAV (tekstitykset)**	**PAV (tekstitykset ja kuvaus)**	**nn_ensemble (tekstitykset)**	**nn_ensemble (tekstitykset ja kuvaus)**
somalit suomen-ruotsalaiset suomalaiset muslimit perhekäsitykset ruokaperinne pukeutuminen kulttuurierot ylpeys ennakkoluulot rasismi ihonväri paluumuuttajat maahanmuuttajat yhteisöt heimot valokuvaajat valokuvanäyttelyt	mielikuvat hyllyt syöminen Somalia Suomi	rasismi mielikuvat vapaaehtoistyö Somalia somalit	Somalia somalit yhteiskunta Suomi Poro	Somalia somalit yhteiskunta Suomi poro

Taulukko 7. Mun heimo -ohjelman asiasanoitus
Alinäyte: 2, ID: PROG_2008_00028879
Yle TV1 , 4.10.2009 klo 19.15-19.45
https://www.rtva.kavi.fi/program/details/program/3970235

Johtopäätöksiä ja ehdotuksia

Kun verrataan koneen tuottamaa asiasanoitusta ihmisen tekemään, kannattaa muistaa että eri yksilöt asiasanoittavat samaa aineistoa erilaisin painoituksin. Keskimäärin noin kolmasosa annetuista asiasanoista on samoja (ks. esim. Sinkkilä & al. 2011, 11). Aiheesta on tehty runsaasti tutkimusta (ks. esim Ingwersen & al., 2005, 42). Kokemattomien ja kokeneiden asiasanoittajien yhdenmukaisuutta asiasanoittamisessa on tutkittu ja havaittu, että yhdenmukaisuuteen voivat vaikuttaa myös käytettävät sanastot ja niiden rakenteet. Eri kokemustason omaavien asiasanoittajien erot asiasanoituksen yhdenmukaisuudessa eivät välttämättä ole kovin suuria (ks. esim. Soler Monreal & al., 2011). Automaattisen asiasanoituksen validointitestissä päästiin parhaimmillaan yli 0,3 F1-lukemaan.

Taustaohjelmat eivät osaa erotella, mikä televisio-ohjelman tekstityksessä on kokonaisuuden kannalta relevanttia. Toisinaan tarjottu käsite saattaa olla täysin epärelevantti. Taustaohjelmia koulutettaessa on syytä tarkkailla myös asiasanoituksen laatua, vaikka validointilukemat näyttäisivätkin hyviltä. Automaattisen asiasanoituksen sanoitustestauksessa nähtiin, kuinka hyvin taustaohjelmat toimivat erityyppisten ohjelmien sanoituksessa. Televisio-ohjelma saattaa olla tehty niin, ettei ohjelman keskeistä aihetta tai teemaa mainita lainkaan, vaan ne esitetään visualisoinnin, musiikin tai äänten avulla. Kun aineistoon sisällytettiin lyhytkin ohjelmakuvaus, tulokset paranivat.

Koulutusaineistoina käytettiin Finnasta poimittua materiaalia sekä ohjelmatekstityksiä ja ohjelmakuvauksia. Epärelevanttien termien havaittiin tulevan pääsääntöisesti Finna-koulutusaineistosta. Automaattisen asiasanoituksen tulokset paranivat, kun koulutusaineistoon lisättiin Ritva-tietokannan sisältöjä. On todennäköistä että tulokset paranisivat entisestään, jos koulutusaineistona olisi pelkästään televisio-ohjelmiin liittyvää aineistoa.

Tulevissa Annifin versioissa asiasanoittajat voisivat parantaa Annifin nn_ensemble-taustaohjelman tuottaman asiasanoituksen laatua itse. Käyttäjät voisivat tarkistaa ehdotetut asiasanat ja lähettää korjatut versiot taustaohjelmalle. Palautteen avulla taustaohjelma voisi muuttaa asiasanoituksen pisteytystä. Voidaan olettaa, että samantyyppinen tv-ohjelma saisi paremmat asiasanaehdotukset jatkossa.

Annifin antamia asiasanoja voi käyttää tarkistetuista asiasanoista erillisinä. Loppukäyttäjille pitää kertoa selvästi, että kyseessä ovat taustaohjelman tuottamat asiasanat.

Jyväskylän yliopiston kirjaston JYX-tietokannassa Annif on otettu käyttöön vuonna 2018. Kun opiskelija julkaisee gradunsa verkossa JYXissä, hän näkee lomakkeella Annifin ehdotukset asiasanoiksi. Opiskelija voi hyväksyä tai

hylätä näitä, ja hänen on mahdollista lisätä omia asiasanojaan. Informaatikko tarkastaa asiasanoituksen ennen julkaisua. Hyväksytyistä ja hylätyistä termeistä pidetään tilastoa. (Häyrinen, 2019.) Jos automaattinen asiasanoitus toteutetaan esimerkiksi Ritvaan, toiminnon käyttöä tulisi tilastoida ja analysoida kuten Jyväskylän tapauksessa. Hyväksyttyjen ja hylättyjen asiasanojen osuuksista on hyödyllistä kerätä tilastotietoa, ja myös seurattava ja analysoitava, kuinka Annifin tuottamia asiasanoja käytetään tiedonhaussa.

Automaattisen asiasanoituksen kokeilu Ylen asiaohjelmien kanssa tuotti niin lupaavia tuloksia, että Annifin käyttöönottoa Ritvassa harkitaan vakavasti. Seuraavaksi testaamme MTV:n fiktiivisten ohjelmien automaattista asiasanoitusta keväällä 2020. Jatkossa yksi mahdollinen käyttötapa voisi olla radio-ohjelmien automaattinen sisällönkuvailu. Kun saadaan muunnettua puhetta tekstimuotoon puheentunnistuksen avulla, voidaan Annifia käyttää vastaavasti automaattisessa asiasanoituksessa.

Lähdeluettelo

Asiasanastot ja ontologiat (2020). Kansalliskirjasto. Saatavilla: https://www.kiwi.fi/display/Asiasanastotjaontologiat/Yleistietoa+ontologioista [viitattu 11.2.2020]

Facebook Research (2019). *fastText/docs/supervised-tutorial.md* . Saatavilla: https://github.com/facebookresearch/fastText/blob/master/docs/supervised-tutorial.md [viitattu 22.10.2019]

Finto (2020a). Kansalliskirjasto. Saatavilla: https://finto.fi/yso/fi/ [viitattu 11.2.2020]

Finto (2020b). Kansalliskirjasto. Saatavilla: https://finto.fi/koko/fi/ [viitattu 11.2.2020]

Harju, E., Kataja, J. & Sainio, T. (2018). *Kansallinen digitaalinen kirjasto; Loppuraportti hankekaudelta 2014–2017.* Helsinki: Opetus- ja kulttuuriministeriö. Noudettu osoitteesta http://urn.fi/URN:ISBN:978-952-263-560-0

Häyrinen, A. (2019). Annif oikeissa töissä. Miten ANNIFia käytetään JYU:n Avoimen tiedon keskuksessa. Saatavilla: https://www.kiwi.fi/pages/viewpage.action?pageId=132677810&preview=/132677810/138936434/Hayrinen_annif_at_work%5B1%5D.pdf [viitattu 11.11.2019]

Ingwersen, P. & Järvelin, K. (2005). *The turn: integration of information seeking and retrieval in context*. Dordrecht, The Netherlands: Springer.

Joulin, A., Grave, E., Bojanowski, P. & Mikolov, T. (2017). Bag of tricks for efficient textclassification. Teoksessa *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Noudettu osoitteesta http://aclweb.org/anthology/E17-2068

Järvelin, K. & Kekäläinen, J. (2000). Kuinka evaluoida tiedonhakumenetelmiä parhaiden dokumenttien löytämisen kannalta? *Informaatiotutkimus, 19*(3), 63–73. Noudettu osoitteesta https://journal.fi/inf/article/view/1612

Kansallinen audiovisuaalinen instituutti (2019a). Saatavilla: http://www.kavi.fi [viitattu 28.10.2019]

Kansallinen audiovisuaalinen instituutti (2019b). RITVA : RTVA:n katselu- ja kuuntelupisteet. Saatavilla: https://www.rtva.kavi.fi/cms/page/page/info_katselupisteet [viitattu 28.10.2019]

Koskimies, A. (2017). *Kielen tunnistus koneoppimismenetelmällä*. Kandidaatintyö. Tampereen teknillinen yliopisto, signaalinkäsittelynlaitos. Noudettu osoitteesta http://URN.fi/URN:NBN:fi:tty-201712142345

Laki sähköisen viestinnän palveluista 7.11.2014/917. Finlex. Saatavilla: https://www.finlex.fi/fi/laki/ajantasa/2014/20140917 [viitattu 11.2.2020].

Leuhu, T. (2014). *Sentiment analysis using machine learning*. Diplomityö. Tampereen teknillinen yliopisto, Signaalinkäsittelyn ja tietoliikennetekniikan koulutusohjelma. Noudettu osoitteesta http://urn.fi/URN:NBN:fi:tty-201505201399

Medelyan, O. (2009). *Human-competitive automatic topic indexing*. University of Waikato. Noudettu osoitteesta https://hdl.handle.net/10289/3513

Pedregosa, F. (2013). *Isotonic Regression*. Saatavilla: http://fa.bianp.net/blog/2013/isotonic-regression/ [viitattu 20.10.2019]

Pouliquen, B., Steinberger, R. & Ignat, C. (2003). Automatic annotation of multilingual textcollections with a conceptual thesaurus. Teoksessa *Proceedings of the Workshop on Ontologies and Information Extraction at the EUROLAN Conference*, Cluj-Napoca, Romania, 19–28. Noudettu osoitteesta https://arxiv.org/abs/cs/0609059

Sanastokeskus ry. (2011). Saatavilla: http://www.tsk.fi/tsk/fi/node/267?page=get_id&id=ID40&vocabulary_code=TSKTT [viitattu 11.2.2020]

Sinkkilä, R., Suominen, O. & Hyvönen, E. (2011). Automatic Semantic Subject Indexing of Web Documents in Highly Inflected Languages. Teoksessa *The Semantic Web: Research and Applications : 8th Extended Semantic Web Conference, ESWC 2011, Heraklion, Crete, Greece, May 29–June 2, 2011, Proceedings*, 215–229. Noudettu osoitteesta https://doi.org/10.1007/978-3-642-21034-1_15

Soler Monreal, M.C. & Gil-Leiva, I. (2011). Evaluation of controlled vocabularies by inter-indexer consistency. *Information Research, 16*(4), paper 502. Saatavilla: http://InformationR.net/ir/16-4/paper502.html [viitattu 17.2.2020]

Suominen, O. (2019a). Annif: DIY automated subject indexing using multiple algorithms. Noudettu osoitteesta http://urn.fi/URN:NBN:fi-fe2019052316853

Suominen, O. (2019b). Backend: nn_ensemble. Saatavilla: https://github.com/NatLibFi/Annif/wiki/Backend%3A-nn_ensemble [viitattu 7. 11 2019]

Suominen, O. (2019c). Annif. Saatavilla: http://annif.org/ [viitattu 30. 10 2019]

Suominen, O. (2019d). Henkilökohtainen tiedonanto Kansalliskirjaston ylläpitämällä #tekoäly-yhteistyö Slack-kanavalla 8.8.2019.

Toepfer, M. & Seifert, C. (2018). Fusion architectures for automatic subject indexing underconcept drift. *International Journal on Digital Libraries*, 1–21. Saatavilla: https://research.utwente.nl/files/80439235/Toepfer2018_ijdl_subject_indexing_under_concept_drift_preprint.pdf [viitattu 30.10.2019]

Traficom. (1. 3 2019). Ääni- ja tekstitysvelvoite televisio-ohjelmissa. Saatavilla: https://www.traficom.fi/fi/viestinta/tv-ja-radio/aani-ja-tekstitysvelvoite-televisio-ohjelmissa [viitattu 21. 10 2019]

Wahlroos, M. (2013). *Indeksointimetatiedon eristäminen ja arviointi*. Pro gradu. Helsingin yliopisto, tietojenkäsittelytieteen laitos. Noudettu osoitteesta http://urn.fi/URN:NBN:fi-fe2017112251247