PROJEKTISUUNNITELMA
SYYSKUU 2019

Automaattisen kuvailun yhteistyöprojekti

Sisällysluettelo

Projektin kuvaus

Aikataulu

Tavoitteet

Osallistujat

Työskentelytavat

Työsisältö
- Työpaketti 1. Julkiset koulutus- ja arviointiaineistot
- Työpaketti 2. Rajatun pääsyn koulutus- ja arviointiaineistot
- Työpaketti 3. Menetelmien vertailu
- Työpaketti 4. Annif-työkalun jatkokehittäminen
- Työpaketti 5. Automaattisen kuvailun työvuot ja tuotannollistaminen
- Työpaketti 6. Viestintä

Työpakettien aikataulutus

Projektin kuvaus

CSC:n ja Kansalliskirjaston automaattisen kuvailun yhteistyöprojektissa kehitetään koneoppimiseen perustuvia automaattisen kuvailun menetelmiä ja työkaluja, sekä edistetään kuvailun automatisaatiota kulttuuriperintöorganisaatioissa.

CSC:n osalta yhteistyö liittyy CEF-rahoitteiseen High-Performance Digitisation (HPD) -hankkeeseen (2017-FI-IA-0132), joka on käynnissä syyskuusta 2018 elokuuhun 2020 ja jonka vastuuhenkilö CSC:llä on Aleksi Kallio. Kansalliskirjaston osalta projekti linkittyy vahvasti Kansalliskirjaston omaan automaattisen kuvailun ja Annif-työkalun kehitykseen. Työtä tehdään osana CSC:n, Kansalliskirjaston ja Kansallisarkiston tekoälykumppanuutta.

Aikataulu

Yhteistyöprojekti käynnistyy syyskuun 2019 alussa ja jatkuu projektimuotoisena vuoden ajan. Tällöin projekti saatetaan päätökseen elokuun lopussa 2020, jonka on myös HPD-hankkeen päättymisaika. Yhteistyön jatkosta päätetään projektin kokemusten pohjalta.

Tavoitteet

Projektin tavoitteena on kehittää ratkaisuja kulttuuriperintöorganisaatioiden digitaalisten tekstiaineistojen automaattiseen kuvailuun. Projektissa testataan ja arvioidaan automaattisen sisällönkuvailun algoritmien soveltuvuutta eri tyyppisille tekstiaineistoille, kehitetään automatisoinnin työkaluja sekä luodaan palvelumalleja automaattisen kuvailun hyödyntämiseksi osana olemassaolevia kuvailuprosesseja.

Yhtenä keskeisenä tavoitteena on Kansalliskirjaston Annif-kuvailutyökalun edelleenkehittäminen CSC:n suurteholaskentaympäristöä hyödyntäen. Projektissa kehitettävät työkalut ja mallit tarjotaan avoimesti kaikkien halukkaiden käyttöön niin kansallisesti kuin kansainvälisestikin. Projektissa tuotettava lähdekoodi on avointa ja aineistot sekä käytetyt kuvailusanastot monikielisiä. Tuotetut ohjelmistot, dokumentoidut kokeilut sekä kerätyt esim. tekoälyn kouluttamiseen soveltuvat aineistot tukevat suomalaisen automaattisen kuvailun yhteisön syntymistä ja toivottavasti laskevat kynnystä uusien menetelmien käyttöönottoon.

Projekti on luonteva osa suomalaisen yhteiskunnan laajempaa digitalisaatiokehitystä ja tukee valtiovallan tavoitteita tekoälyn käytön lisäämisestä julkisessa hallinnossa.

Osallistujat

Projektiin osallistuvat Kansalliskirjaston ja CSC:n asiantuntijat, jotka työskentelevät automaattisen kuvailun parissa.

Kansalliskirjaston osalta projektiin osallistuvat:

- Osma Suominen, Annif-kehittäjä
- Juho Inkinen, Annif-kehittäjä
- Mona Lehtinen, Annif-kehittäjä
- Satu Niininen, sisällönkuvailun ja e-kuvailuaineiston asiantuntija
- Mikko Lappalainen, koordinaattori

Yhteyshenkilönä toimii Mikko Lappalainen.

CSC:n osalta projektiin osallistuvat:

- Markus Koskela, koneoppimisen asiantuntija
- Mats Sjöberg, koneoppimisen asiantuntija
- Juha Hulkkonen, data engineer
- Aleksi Kallio, koordinaattori

Yhteyshenkilönä toimii Aleksi Kallio.

Projektin etenemistä seuraa kvartaaleittan kokoontuva, projektiryhmää laajempi yhteistyöhankkeen seurantaryhmä. Seurantaryhmä hyväksyy kunkin kvartaalin päätteeksi valmistuneet työpaketit, sekä keskustelee jatkotyöskentelyn päälinjoista.

Työskentelytavat

Yhteistyöprojektissa pyritään ketterään työskentelytapaan. Projektin osallistujat kokoontuvat säännöllisesti etäkokoukseen, jossa todetaan työpakettien eteneminen ja sovitaan seuraavista tehtävistä. Etäkokoukset järjestetään videoneuvotteluna kahden viikon välein ja niistä tehdään muistiinpanot jaettuun dokumenttikansioon (https://tinyurl.com/akyjako).

Muussa projektin kommunikaatiossa ensisijainen kanava on Kansalliskirjaston Slack-kanava #tekoäly-yhteistyö. Työtehtäviä hallitaan Trellon kautta (https://trello.com/b/rGZRHxQn/automaattisen-kuvailun-yhteisty%C3%B6projekti).

Työsisältö

Projekti koostuu seuraavista työpaketeista sekä työpakettien alaisista tehtävistä:

Työpaketti 1. Julkiset koulutus- ja arviointiaineistot

Kuvaus: Kerätään julkisesti jaettavia koulutusaineistoja automaattisen kuvailun koneoppimista varten. Aineistot julkaistaan GitHubin julkisessa Annif-corpora -varastossa. Suurimmaksi osaksi nämä aineistot on jo julkaistu ennen projektin alkamista, mutta ne on kirjattu työtehtäviksi, jotta projektissa käytettävistä aineistoista muodostuisi kokonaiskuva.

Tehtävä 1.1. Finna-pohjainen YSO-koulutusaineisto
Algoritmien kouluttamiseen tarkoitettu kolmikielinen (fi, sv, en) aineisto, joka on muodostettu Finna.fi-hakupalvelun viitetietueista, joilla on Yleisestä suomalaisesta ontologiasta (YSO) poimitut asiasanat. Julkaistu 6/2019. Korvaa aiemman vuonna 2017 koostetun Finna-pohjaisen koulutusaineiston, jolla Annifin kehitys aloitettiin.
Tietuemäärät: fi 6,6M; en 2,0M; sv 0,7M.
Aineisto on jo valmis ja julkaistu Annif-corpora-varastossa.

Tehtävä 1.2. Finna-pohjainen YKL-koulutusaineisto
Algoritmien kouluttamiseen tarkoitettu kolmikielinen (fi, sv, en) aineisto, joka on muodostettu Finnan viitetietueista, joilla on Yleisten kirjastojen luokitusjärjestelmästä (YKL) poimitut luokat.
Päävastuu: KK.
Tuotokset: Aineisto julkaistuna Annif-corpora-varastossa.

Tehtävä 1.3. Kysy kirjastonhoitajalta -aineisto
Kokotekstiaineisto, joka on muodostettu Kysy kirjastonhoitajalta -palvelun suomenkielisistä kysymys-vastauspareista. Julkaistu 4/2018. Yhteensä 3150 dokumenttia. Aineistoa voidaan käyttää sekä algoritmien kouluttamiseen että arviointiin.
Aineisto on jo valmis ja julkaistu Annif-corpora-varastossa.

Tehtävä 1.4. Jyväskylän yliopiston opinnäytteet
Kokotekstiaineisto, joka on muodostettu Jyväskylän yliopiston julkaisuarkistosta poimituista suomen-, ruotsin- ja englanninkielisistä opinnäytteistä (gradut ja väitöskirjat). Julkaistu 4/2018. Yhteensä 7400 dokumenttia. Aineistoa voidaan käyttää sekä algoritmien kouluttamiseen että arviointiin.
Aineisto on jo valmis ja julkaistu Annif-corpora-varastossa.

Työpaketti 2. Rajatun pääsyn koulutus- ja arviointiaineistot

Kuvaus: Kerätään koulutusaineistoja automaattisen kuvailun koneoppimista varten, jotka voidaan jakaa projektin osapuolten kesken.

Tehtävä 2.1. Vapaakappaleluovutuksiin perustuva arviointiaineisto (YSO)
Aineisto perustuu Kansalliskirjaston e-vapaakappalejärjestelmään (Varia) vuosina 2013-2019 kerättyihin e-kirjoihin, jotka ovat suomen-, ruotsin- tai englanninkielisiä, joilla on ISBN-tunniste ja joista löytyy YSO-sisällönkuvailu Kansallisbibliografiasta (Fennica). Kokotekstiteoksia on yhteensä noin 9600 kpl ja raakatekstiä noin 2 Gt. Tekijänoikeussyistä aineiston sanajärjestys sekoitetaan ennen luovutusta CSC:lle. Aineistoa voidaan käyttää sekä algoritmien kouluttamiseen että arviointiin.
Päävastuu: KK
Tuotokset: Aineisto luovutettuna CSC:lle esim. Funet Filesender-palvelun kautta.

Tehtävä 2.2. Kirjaesittelyihin perustuva arviointiaineistot (YSO)
Aineisto perustuu Kirjavälitys Oy:lta saatuihin kirjojen suomen-, ruotsin- ja englanninkielisiin esittelyteksteihin sekä ko. kirjojen YSO-sisällönkuvailuun Kansallisbibliografiassa (Fennica). Aineistoa voidaan käyttää sekä algoritmien kouluttamiseen että arviointiin.
Päävastuu: KK
Tuotokset: Aineisto luovutettuna CSC:lle esim. Funet Filesender-palvelun kautta.

Tehtävä 2.3. Kirjaesittelyihin perustuva koulutus- ja arviointiaineisto (YKL)

Aineisto perustuu Kirjavälitys Oy:lta saatuihin kirjojen suomen-, ruotsin- ja englanninkielisiin esittelyteksteihin sekä YKL-luokkiin. Aineistoa voidaan käyttää sekä algoritmien kouluttamiseen että arviointiin.
Päävastuu: KK
Tuotokset: Aineisto luovutettuna CSC:lle esim. Funet Filesender-palvelun kautta.

Työpaketti 3. Menetelmien vertailu

Kuvaus: Vertaillaan erilaisia koneoppimisen menetelmiä automaattisen kuvailun tarpeita ajatellen ja tutkitaan, että kuinka hyvin ne toimivat erityyppisillä aineistolla ja millaisia virheitä eri menetelmät tekevät. Vertailussa hyödynnetään CSC:n suurteholaskennan ympäristöä ja laajasti erilaisia aineistoja.

Tehtävä 3.1. FastText-hyperparametrien etsiminen YSO- ja YKL-koulutusaineistoille
FastText-algoritmi on jo havaittu hyväksi ja integroitu Annifiin, mutta sen käyttö edellyttää toimivien hyperparametrien löytämistä. Hyperparametrien etsintä vaatii raskasta laskentaa. Tavoitteena on löytää toimivat parametrit työpakettien 1 ja 2 aineistojen käyttöön.
Päävastuu: CSC
Tuotokset: Suositus toimivista hyperparametreista eri käyttötarkoituksiin (YSO, YKL, mahdollisesti jaoteltuna käytettyjen koulutus- ja/tai arviointiaineistojen mukaan)

Tehtävä 3.2. Koneoppimismenetelmien testaus
Tavoitteena on saada selville miten hyvin erilaiset olemassaolevat perinteiset, ei-neuroverkkopohjaiset koneoppimisalgoritmit (esim. FastXML-variantit, Parabel, PDSparse, AnnexML, sklearn-hierarchical-classification) soveltuvat tekstidokumenttien automaattiseen sisällönkuvailuun eri koulutus- ja arviointiaineistoilla.
Päävastuu: CSC
Tuotokset: Taulukko testituloksista eri koulutus- ja arviointiaineistoille

Tehtävä 3.3. Neuroverkkomenetelmien mahdollisuuksien tutkiminen
Modernien neuroverkkoarkkitehtuurien ja -menetelmien soveltuvuuden selvittäminen yhteistyöprojektin aihepiiriin liittyviin teemoihin. Viime vuosina on julkaistu useita tekstiaineistojen analysointiin soveltuvia kehittyneitä neuroverkkomenetelmiä. Tässä tehtävässä arvioidaan ja pilotoidaan näiden soveltuvuutta sisällönkuvailuun ja siihen läheisesti liittyviin sovelluskohteisiin.
Päävastuu: CSC
Tuotokset: Arvio menetelmien käyttökelpoisuudesta ja jatkoselvityksen aiheista

Tehtävä 3.4. Suositus toimiviksi havaituista algoritmeista
Tehtävissä 3.2 ja 3.3 suoritettujen testien perusteella suositellaan 1-3 algoritmia (ja niiden mahdollisia hyperparametreja), jotka kannattaisi integroida Annifiin.
Päävastuu: CSC

Tuotokset: Johtopäätös sekä suositus tehtyjen testien perusteella, sekä mahdollisesti lyhyt opasmuotoinen teksti menetelmien valitsemisesta ja käyttöönotosta automaattisen kuvailun tarpeisiin

Työpaketti 4. Annif-työkalun jatkokehittäminen

Kuvaus: Hyväksi havaitut menetelmät viedään osaksi Annif-työkalua.
Päävastuu: KK
Tuotokset: 1-3 uutta Annifiin integroitua algoritmia

Tehtävä 4.1. Testattujen FastText-hyperparametrien käyttöönotto Annifissa
Viedään tehtävässä 3.1 löydetyt fastText-hyperparametrit käytäntöön Annifin esimerkkikonfiguraatioissa, dokumentaatiossa ja asennuksissa
Päävastuu: KK
Tuotokset: Annifin konfiguraatiot ja dokumentaatiot muutettu

Tehtävä 4.2. Uusien algoritmien integrointi Annifiin
Integroidaan tehtävässä 3.4 suositellut algoritmit Annifiin.
Päävastuu: KK
Tuotokset: Uusia backend-moduuleja Annifin koodissa sekä niiden dokumentointi

Työpaketti 5. Automaattisen kuvailun työvuot ja tuotannollistaminen

Kuvaus: Tämä työpaketti tuotannollistaa koneoppimisen menetelmiä käytettäväksi osana Kansalliskirjaston ja muiden muistiorganisaatioiden toimintaa. Koneoppiminen ja tekoälyjärjestelmien rakentaminen tuovat omat uudenlaiset haasteensa ohjelmistokehitykseen. Työpaketissa jaetaan kokemuksia, osaamista ja koodia automaattisen kuvailun työvoihin liittyen; tutkitaan Dockerin, Kuberneteksen ja muiden keskeisten teknologioiden soveltuvuutta työvoiden paketointiin, jakeluun ja ajamiseen; sekä dokumentoidaan löydökset.

Tehtävä 5.1. Ongelmakentän kuvaus
Kuvataan automaattisen kuvailun työvuota hajautetun palvelinjärjestelmän näkökulmasta. Millaisia ratkaisuja tarvitaan, jotta voidaan tuottaa jatkuvaa teknistä palvelua, jossa on mukana koneoppimiseen perustuvia komponentteja? Miten opittuja malleja päivitetään datan määrän kasvaessa? Miten suojaudutaan mahdollisilta vihamielisiltä käyttäjiltä? Kuinka varmistetaan järjestelmän skaalautuvuus käytön tai datan määrän kasvaessa?

Päävastuu: CSC
Tuotokset: Lyhyt artikkeli (white paper) automaattisen kuvailun järjestelmän suunnittelussa huomioitavista asioista

Tehtävä 5.2. PoC automaattisen kuvailun palvelun integroimisesta
Paketissa laaditaan proof of concept -selvitys Annifin integroimisesta Kansalliskirjaston olemassaoleviin kuvailuprosesseihin. Selvitys sisältää nykyisten prosessien kuvauksen, ehdotuksen Annifin roolista osana prosesseja, sekä kuvauksen Annifin käytännön integroimisesta kuvailuympäristöön. Lisäksi selvitys kuvaa palvelun parhaat päivitysprosessit, mallien ja sanastojen muutostenhallinnan ja käyttöoikeuksien hallinnan. Tämä PoC toimii jatkossa alustavana työsuunnitelmana käytännön käyttöönottotyölle.
Päävastuu: KK
Tuotokset: Proof of concept -selvitys.

Tehtävä 5.3. Infrastruktuurin paketointi
Tuotetaan pilottiluontoisesti kompakti paketointi, jolla realistinen tai realistisuuteen pyrkivä automaattisen kuvailun palvelinympäristö voidaan ottaa käyttöön. Tavoitteena on madaltaa kynnystä automaattisen kuvailun käyttöönottoon sekä yhtenäistää käytettyjä teknologioita tarjoamalla valmiiksi mietitty kokonaisuus. Pyritään tekemään paketointi, joka on käytettävissä sekä raskaassa palvelinympäristössä, että toisaalta yksittäisen kehittäjän läppärillä, esim. hyödyntäen Kubernetes-teknologiaa.
Päävastuu: CSC
Tuotokset: Käyttövalmis template sekä lyhyt ohjeistus

Työpaketti 6. Viestintä

Kuvaus: Viestitään yhteistyöstä ja projektin tuloksista laajasti niin kotimaassa kuin kansainvälisestikin. Viestinnässä hyödynnetään CSC:n ja Kansalliskirjaston olemassaolevia viestintäkanavia.

Päävastuu: CSC ja KK
Tuotokset:

Työpakettien aikataulutus

	9-11/2019	**12/2019-2/2020**	**3-5/2020**	**6-8/2020**
T1.1. Finna-YSO korpus	valmis			
T1.2. Finna-YKL korpus				
T1.3. Kirjastonhoitaja-korpus	valmis			
T1.4. JYU korpus	valmis			
T2.1. Vapaakappale-korpus				
T2.2. Kirjaesittely-YSO korpus				
T2.3. Kirjaesittely-YKL korpus				
T3.1. fastText parametrien haku				

T3.2. Koneoppimisen testaus				
T3.3. Neuroverkkotestaus				
T3.4. Suositus algoritmeista				
T4.1. fastText-parametrit käyttöön				
T4.2. Uudet algoritmit Annifiin				
T5.1. Ongelmakentän kuvaus				
T5.2. PoC palvelun integroinnista				
T5.3. Infran paketointi				
T6 Viestintä				
CSC päävastuu				
KK päävastuu				