Metatietojen tunnistus

Metatietojen tunnistus koostuu kolmesta erillisestä osasta: nimientiteettien tunnistus, automaattinen asiasanoitus ja dokumentin kielen tunnistus.

Nimientiteettien tunnistus

Tausta, hyöty ja käyttötarkoitus

Nimientiteettien avulla voidaan yhdistää ja etsiä erilaisiin aiheisiin liittyviä arkistoyksiköitä ja asiakirjoja, jolloin loppukäyttäjän on mahdollista löytää sellaistakin aineistoa, jota ei alun perin olisi osannut etsiä.

Komponentti hyödyntää suomen- tai englanninkielistä nimientiteettien tunnistusmallia dokumentin tekstisisällön automaattisesti tunnistetun kielen perusteella. Suomenkielinen malli rakentuu BERT-kielimallin suomalaisen version (https://github.com/TurkuNLP/FinBERT) pohjalle ja on koulutettu ja kehitetty itse, kun taas englanninkielinen nimientiteettien tunnistus perustuu Spacy-kirjaston (https://spacy.io/models/en#en_core_web_trf) tarjoamaan sovellukseen. Suomenkielisen mallin tunnistamia entiteettiluokkia ovat

- **Päivämäärät**: esim. 1.10.2016, 1. lokakuuta, vuoden 2016 aikana, 1980-luvulla
- **Organisaatiot**: esim. Apple, Turun yliopisto, Kokoomus, Keskusrikospoliisi
- **Henkilönimet**: esim. Sauli Niinistö, Joulupukki, @digikim
- **Diaarinumerot**: esim. VNK/123/45/1999, 1/23/56, 1000-1100/123/1988
- **Geopoliittiset paikannimet**: esim. Päijät-Häme, Helsinki, HKI, Katajanokka, Lappi
- **Muut (ei-geopol.) paikanimet**: esim. Yosemiten kansallispuisto, Mars, Atlantti, Kemijoki
- **Tuotteet**: esim. IPhone 6, C++, Helsingin Sanomat, Patriot Act -laki
- **Tapahtumat**: esim. Mobile World-tapahtuma, Toinen maailmansota, Covid-19
- **Y-tunnukset**: kts. https://www.ytj.fi/index/y-tunnus.html
- **Kansallisuudet, uskonnolliset ja poliittiset ryhmät**: esim. suomalaiset, suomalaisuus, suomenkielinen, muslimit, Elokapina

Englanninkielinen malli tunnistaa diaarinumeroita ja y-tunnuksia lukuun ottamatta samat luokat, joskin luokkien sisällön määrittely ei kaikissa tapauksissa vastaa täysin yllä esitettyä.

Mitä komponentti tekee?

Komponentti saa syötteenä konekirjoitetun tekstin tunnistuksen (OCR) avulla digitoidusta dokumentista tunnistetun tekstisisällön, josta se etsii yllä listattuihin luokkiin kuuluvia metatietoja. Komponentti palauttaa luokkakohtaisen listauksen dokumentista löydetyistä metatiedoista, joista on poistettu mahdolliset täysin identtisessä muodossa toistuvat ilmentymät (esim. vaikka sana ’Helsinki’ toistuisi tekstissä useamman kerran, tuloksissa se esiintyy vain kerran).

Miten komponenttia käytetään?

Komponentti on saatavilla sekä osana Arkkiivi-käyttöliittymää (http://www.arkkiivi.fi/) että itsenäisenä, ohjelmointirajapinnan (API) tarjoavana sovelluksena DALAI-hankkeen GitHub-sivulla (https://github.com/DALAI-project/NER_API). Mallitiedosto on ladattavissa HuggingFace-sivustolta (https://huggingface.co/Kansallisarkisto/finbert-ner). Tarkemmat ohjeet komponentin käyttöön eri ympäristöissä löytyvät oheisilta verkkosivuilta.

Komponentin koulutus

Komponentin koulutukseen on käytetty sekä Turku OntoNotes Entities Corpus-aineistoa (https://github.com/TurkuNLP/turku-one), NewsEye-aineiston (https://zenodo.org/record/4694466#.YJR20qE6-bi) suomenkielistä osaa, että Kansallisarkiston digitoimista asiakirjoista koostettua ja annotoitua aineistoa. Itse tuotettu koulutusaineisto on ensin ajettu konekirjoitetun tekstin tunnistuksen (OCR) läpi, jonka jälkeen tekstistä on manuaalisesti annotoitu yllä lueteltuihin kategorioihin kuuluvia nimientiteettejä.

Kokonaisuudessaan komponentin koulutusdata sisältää n. 105 000 entiteettiä. Koulutuksessa käytetyn aineiston määrä ja ajallinen kattavuus (1800-luvun puolivälistä nykypäivään) on rajallinen, mikä tekstintunnistuksen laadun ohella osaltaan vaikuttaa siihen, että komponentin tunnistuksessa tapahtuu myös virheitä.

Asiasanoitus

Tausta, hyöty ja käyttötarkoitus

Komponentti pohjaa Kansalliskirjaston kehittämään Annif-ohjelmistoon, joka on jatkuvan ja aktiivisen kehityksen kohteena. Voit käydä tutustumassa aiheeseen osoitteessa https://annif.org/. Asiasanoitusta käytetään tiedon etsimiseen aihetunnisteiden perusteella ilman, että koko dokumenttia tarvitsee käydä läpi. Näin esimerkiksi valitun asiasanan sisältämät dokumentit löydetään tehokkaasti. Alun perin Annif kehitettiin erilaisten lopputöiden ja tieteellisten artikkelien asiasanoittamiseen, mutta tulokset osoittavat, että se on varsin yleiskäyttöinen. Annif-ohjelmistoa käyttävät esim. Yleisradio ja Saksan kansalliskirjasto.

Mitä komponentti tekee?

Komponentti kokoaa yhteen erilaisia asiasanoitusmalleja hyödyntäessään Annif-ohjelmistoa, joka pohjautuu koneoppimisen ja kieliteknologian ratkaisuihin. Lopputuloksena on dokumentin kuvaus asiasanoittain.

Tiedostoformaatit ja mahdolliset esikäsittelyt

Komponentti on saatavilla osana Arkkiivi-käyttöliittymää (http://www.arkkiivi.fi/), jossa tuetut tiedostoformaatit ovat tällä hetkellä: .pdf, .jpg, .tif, .tiff, .xml ja .txt. Mikäli tiedosto tulkitaan digisyntyiseksi, sille ei suoriteta konekirjoitetun tekstin tunnistusta (OCR). Mikäli kyseessä on kuvatiedosto, OCR-käsittely suoritetaan Apache Tika -ohjelmiston (https://tika.apache.org/) avulla, sillä Annif osaa tulkita vain tekstisisältöä.

Tulokset

Komponentti palauttaa X-määrän (oletuksena 10) asiasanaa .json-muotoisena datana, jonka voi ottaa käyttöliittymästä ulos (export) myös .csv-muotoisena tiedostona.

Komponentin koulutuksessa käytetty aineisto

Annifin koulutuksessa on käytetty seuraavia sanastoja: YSO-suomi, YSO-english ja ALLFO-svenska.

Kielen tunnistus

Dokumentin kielen tunnistus perustuu tekstin tunnistuksen tuloksena saatavaan dokumentin tekstisisältöön. Kielen tunnistukseen käytetään Apache Tika-ohjelmiston (https://tika.apache.org/) tarkoitusta varten kehitettyä toiminnallisuutta.

Metadata identification

Metadata identification consists of three parts: named entity recognition (NER), automatic subject indexing and document language identification.

Named entity recognition

Background, benefits and purpose of use

With the help of named entities, it is possible to combine and search for archive units and documents related to different topics, so it is possible for the end user to find even material that they would not have been able to search for in the first place.

The component utilises a Finnish or English name-entity identification model based on the automatically recognised language of the document's text content. The Finnish model is built on the Finnish version of the BERT language model (https://github.com/TurkuNLP/FinBERT) and has been trained and developed by the National Archives of Finland in cooperation with the FIN-CLARIAH research infrastructure / University of Jyväskylä (https://www.jyu.fi/hytk/fi/tutkimus/infrastruktuurit/fin-clariah/). The English named entity recognition is based on an application provided by the Spacy library (https://spacy.io/models/en#en_core_web_trf). The entity classes recognised by the Finnish language model are

- Dates: e.g. 1 October 2016, during 2016, in the 1980s
- Organisations: e.g. Apple, University of Turku, National Coalition Party, National Bureau of Investigation
- Personal names: e.g. Sauli Niinistö, Santa Claus, @digikim
- Journal numbers: e.g. NRK/123/45/1999, 1/23/56, 1000-1100/123/1988
- Geopolitical place names: e.g. Päijät-Häme, Helsinki, HKI, Katajanokka, Lapland
- Other (non-geopolitical) place names: e.g. Yosemite National Park, Mars, Atlantic Ocean, Kemijoki
- Products: e.g. IPhone 6, C++, Helsingin Sanomat, Patriot Act
- Events: e.g. Mobile World event, World War II, Covid-19
- Business IDs: see https://www.ytj.fi/index/y-tunnus.html
- Nationalities, religious and political groups: e.g. Finns, Finnishness, Finnish-speaking, Muslims, Elokapina

The English model recognises the same classes except for journal numbers and business IDs, although the definition of the content of the classes does not fully correspond to the above in all cases.

What does the component do?

The component receives as input the text content identified in the digitised document with the help of optical character recognition (OCR), from which it searches for metadata belonging to the classes listed above. The component returns a class-specific list of metadata found in the document, from which any instances occurring in a completely identical form have been removed (e.g., even if the word 'Helsinki' was repeated several times in the text, it only appears once in the results).

How is the component used?

The component is available both as part of the Arkkiivi user interface (http://www.arkkiivi.fi/) and as a standalone application offering an application programming interface (API) on the DALAI project's GitHub page (https://github.com/DALAI-project/NER_API). The model file can be downloaded from the HuggingFace website (https://huggingface.co/Kansallisarkisto/finbert-ner). More detailed instructions on how to use the component in different environments can be found on the attached website.

Component training

Turku OntoNotes Entities Corpus material (https://github.com/TurkuNLP/turku-one) , the Finnish-language part of the NewsEye material (https://zenodo.org/record/4694466#.YJR20qE6-bi) and material compiled and annotated from digitised documents of the National Archives have been used for the training of the component. The self-produced training material has first been run through optical character recognition (OCR), after which name entities belonging to the classes listed above have been manually annotated from the text. In total, the component's training data includes approximately 105,000 entity instances.

The amount and time coverage of the material used in the training (from the mid-19th century to the present) is limited, which in addition to the quality of OCR can lead to errors in the recognition of named entities.

Subject indexing

Background, benefits and purpose of use

The component is based on the Annif software (https://annif.org/) developed by the National Library of Finland, which is under continuous and active development. Index terms are used to search for information based on hashtags without having to go through the entire document. This way documents containing the selected index terms can be found efficiently. Originally, Annif was developed for the subject indexing of various final theses and scientific articles, though the results show that it is quite a general-purpose tool. Annif software is used for example by the Finnish Broadcasting Company and the National Library of Germany.

What does the component do?

The component brings together different subject indexing models when utilising the Annif software, which is based on machine learning and language technology solutions. The end result is a description of the document by keywords.

File formats and possible preprocessing

The component is available as part of the Arkkiivi interface (http://www.arkkiivi.fi/), where the file formats currently supported are: .pdf, .jpg, .tif, .tiff, .xml and .txt. If the file is interpreted as born-digital, no optical character recognition (OCR) will be performed on it. In the case of an image file, the OCR processing is carried out using the Apache Tika software (https://tika.apache.org/) , as Annif is only capable of interpreting text content.

Results

The component returns X number (default 10) of index terms as .json format data, which can also be exported from the interface as a .csv format file.

Material used in component training

The following glossaries have been used in Annif's training: YSO-suomi, YSO-english and ALLFO-svenska.

Language recognition

Document language recognition is based on document text content obtained through optical character recognition. The functionality of the Apache Tika software (https://tika.apache.org/) developed for the purpose is used for language recognition.

Igenkänning av metadata

Igenkänningen av metadata består av tre separata delar: igenkänning av namnentiteter, automatisk ämnesordsindexering och igenkänning av dokumentets språk.

Igenkänning av namnentiteter

Bakgrund, fördelar och ändamål

Med namnentiteter är det möjligt att kombinera och söka efter arkivenheter och dokument som rör olika ämnen. Då kan slutanvändaren även hitta material som hen inte ursprungligen hade förmått söka.

Komponenten använder en finskspråkig eller engelskspråkig modell för igenkänning av namnentiteter baserat på det automatiskt igenkända språket i dokumentets textinnehåll. Den finskspråkiga modellen bygger på den finska versionen av BERT-språkmodellen (https://github.com/TurkuNLP/FinBERT) och har utbildats och utvecklats själv, medan den engelska igenkänningen av namnentiteter bygger på en app som tillhandahålls av Spacy-biblioteket (https://spacy.io/models/en#en_core_web_trf). De entitetskategorier som den finskspråkiga modellen har känt igen är

• datum: t.ex. 1.10.2016, 1 oktober, under 2016, på 1980-talet

• organisationer: t.ex. Apple, Åbo universitet, Samlingspartiet, Centralkriminalpolisen

• personnamn: t.ex. Sauli Niinistö, Julgubben, @digikim

• diarienummer: t.ex. SRK/123/45/1999, 1/23/56, 1000-1100/123/1988

• geopolitiska ortnamn: t.ex. Päijänne-Tavastland, Helsingfors, H:fors, Skatudden, Lappland

• andra ortnamn (ej geopol.): t.ex. Yosemite nationalpark, Mars, Atlanten, Kemi älv

• produkter: t.ex. IPhone 6, C++, Helsingin Sanomat, Patriot Act-lagen

• händelser: t.ex. Mobile World Event, Andra världskriget, covid-19

• FO-nummer: se https://www.ytj.fi/sv/index/y-tunnus.html

• nationaliteter, religiösa och politiska grupper: t.ex. finländare, finländshet, finskspråkiga, muslimer, Extinction Rebellion

Med undantag för diarienummer och FO-nummer känner den engelska modellen igen samma kategorier, fastän definitionen av innehållet i kategorierna inte i alla situationer helt motsvarar det ovannämnda.

Vad gör komponenten?

Med hjälp av igenkänningen av maskinskriven text (OCR) får komponenten ett igenkänt textinnehåll som indata från det digitaliserade dokumentet, där den söker metadata i de kategorier som anges ovan. Komponenten återställer en klasspecifik lista över metadata som hittats i dokumentet och från vilka eventuella upprepade uttryck i helt identisk form har raderats (t.ex. även om ordet Helsingfors upprepas flera gånger i texten, förekommer det endast en gång i resultatet).

Hur används komponenten?

Komponenten finns tillgänglig både som en del av Arkkiivi-gränssnittet (http://www.arkkiivi.fi/) och som en fristående app som tillhandahåller ett programmeringsgränssnitt (API) på DALAI-projektets GitHub-webbplats (https://github.com/DALAI-project/NER_API). Modellfilen kan laddas ner på webbplatsen

HuggingFace (https://huggingface.co/Kansallisarkisto/finbert-ner). Närmare anvisningar om användningen av komponenten inom olika miljöer finns på den webbplatsen.

Utbildning av komponenten

För utbildningen av komponenten användes såväl Turku OntoNotes Entities Corpus-materialet (https://github.com/TurkuNLP/turku-one), den finskspråkiga delen av NewsEye-materialet (https://zenodo.org/record/4694466#.YJR20qE6-bi) och materialet som har sammanställts och annoterats från handlingar som Riksarkivet har digitaliserat. Det självproducerade utbildningsmaterialet har först körts via igenkänningen av maskinskriven text (OCR), varefter namnentiteter som ingår i kategorierna ovan har annoterats från texten manuellt. Totalt innehåller komponentens utbildningsdata cirka 105 000 entiteter.

Mängden och den tidsmässiga omfattningen av det material som använts i utbildningen (från mitten av 1800-talet till i dag) är begränsad, vilket vid sidan av kvaliteten på textigenkänningen bidrar till att det också uppstår fel i komponentens igenkänning.

Ämnesordsindexering

Bakgrund, fördelar och ändamål

Komponenten baserar sig på Annif-programvaran (https://annif.org/) som utvecklats av Nationalbiblioteket och som är föremål för kontinuerlig och aktiv utveckling. Ämnesordsindexeringen används för att söka information utifrån fyrkantstaggar utan att hela dokumentet behöver gås igenom. På så sätt kan man till exempel på ett effektivt sätt hitta dokument som innehåller ett utvalt ämnesord. Annif utvecklades ursprungligen för ämnesordsindexering av olika examensarbeten och vetenskapliga artiklar, men resultaten visar att det kan användas relativt allmänt. Annif-programvaran används av t.ex. Rundradion och Tyska nationalbiblioteket.

Vad gör komponenten?

Komponenten samlar ihop olika modeller för ämnesordsindexering när den använder Annif, som bygger på lösningar för maskininlärning och språkteknik. Slutresultatet är en beskrivning av dokumentet efter ämnesord.

Filformat och eventuella förbehandlingar

Komponenten finns tillgänglig som en del av Arkkiivi-gränssnittet (http://www.arkkiivi.fi/), där följande filformat som stöds för närvarande är: .pdf, .jpg, .tif, .tiff, .xml och .txt. Om en fil tolkas som digitalt skapad, genomgår den inte en igenkänning av maskinskriven text (OCR). Om det är fråga om en bildfil, utförs OCR-behandlingen med programvaran Apache Tika (https://tika.apache.org/), eftersom Annif endast kan tolka textinnehåll.

Resultat

Komponenten återställer X ämnesord (10 som standard) som data i .json-format, som kan exporteras från gränssnittet även som en fil i .csv-format.

Material som använts i utbildningen av komponenten

I utbildningen av Annif har följande ordförråd använts: YSO-suomi, YSO-english och ALLFO-svenska.

Igenkänning av språk

Igenkänningen av språket i dokument bygger på det textinnehåll som textigenkänningen resulterar i. För igenkänning av språk används en funktion som har tagits fram för programvaran Apache Tika (https://tika.apache.org/).