

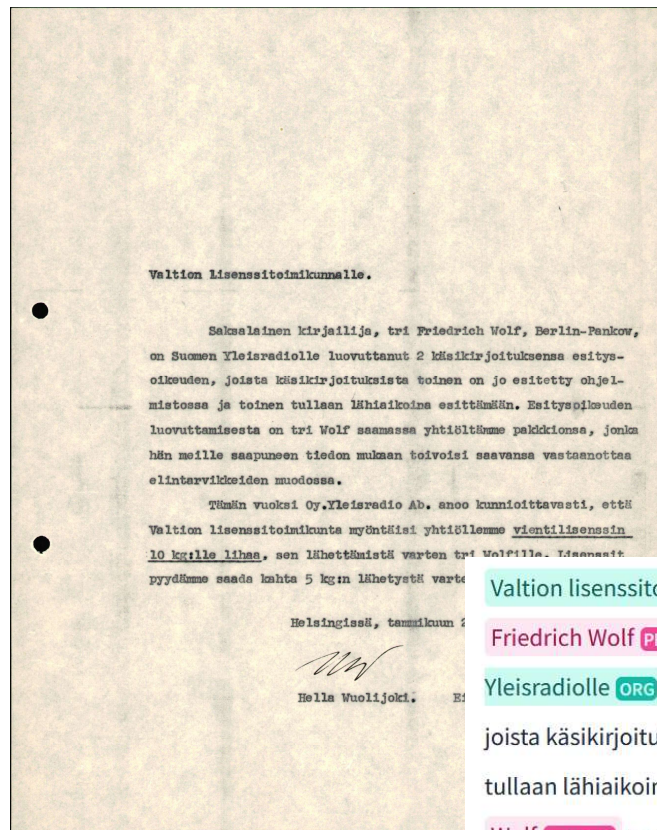


KANSALLISARKISTO

Nimientiteettien tunnistus arkistoaineistosta

Mikko Lipsanen

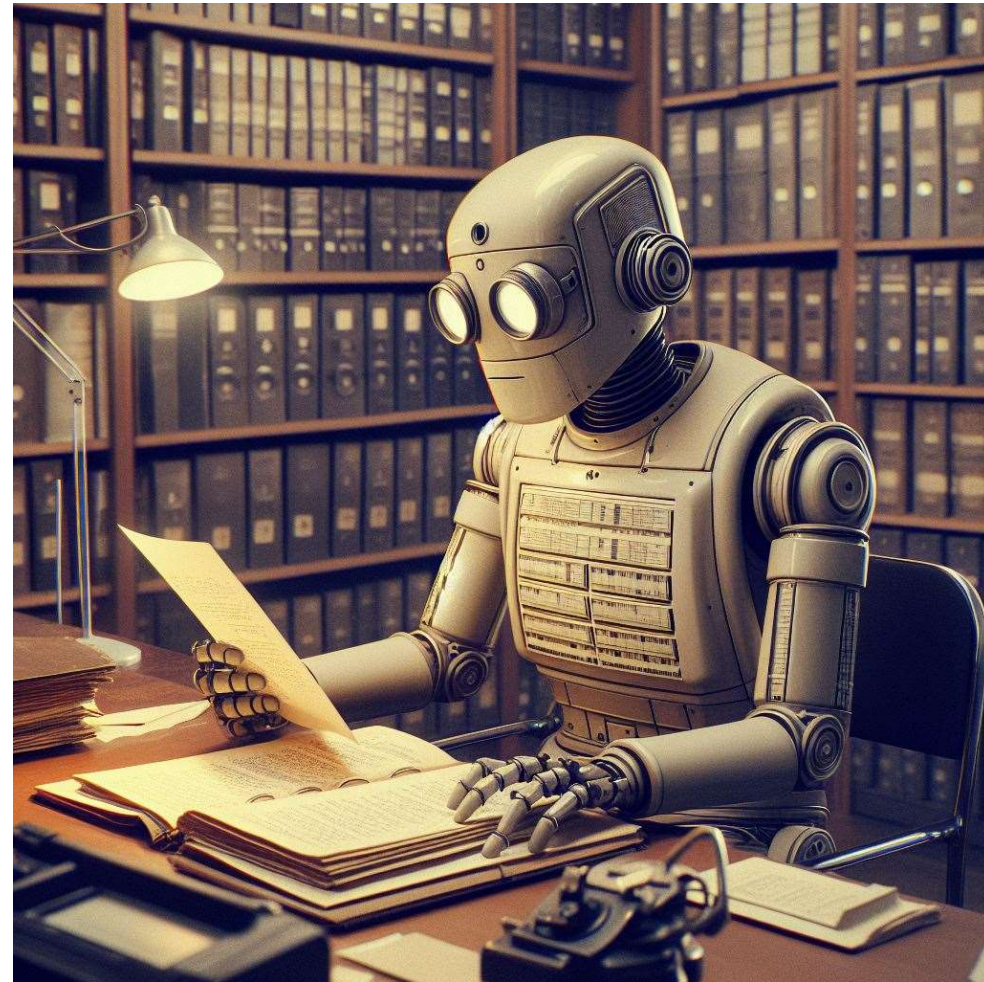
Kirjasto kohtaa tekoälyn - 1.2.2024



Valtion lisenssitoimikunnalle **ORG** , Saksalainen **NORP** Kkirjailija, tri Friedrich Wolf **PERSON**), Berlin-Pankow **PERSON** , on Suomen Yleisradiolle **ORG** luovuttanut 2 käsikirjoituksensa esitys- oikeuden, joista käsikirjoituksista toinen on jo esitetty ohjel- mistossa ja toinen tullaan lähiaikoina esittämään. Esitysoikeuden luovuttamisesta on tri Wolf **PERSON** saamassa yhtiöltämme pakkionsa, jonka hän meille saapuneen tiedon mukaan toivoisi saavansa vastaanottaa elintarvikkeiden muodossa. Tämän vuoksi Oy.Yleisradio Ab. anoo kunnioittavasti, että Valtion lisenssitoimikunta **ORG** myöntäisi yhtiöllemme vientilisenssin 10 kg:lle lihaa, sen lähettämistä varten tri Wolfille **PERSON** . Lisenssit pyydämme saada kahta 5 kg:n lähetyksiä varten. | Helsingissä **GPE** , tammikuun 20 päivä/1948 **DATE** .: AA 7 Hella Wuolijoki **PERSON** . Einar Sundström **PERSON** .

Esityksen sisältö

1. Nimientiteettien tunnistus (NER): Mistä on kyse?
2. Suomenkielinen NER: Mallit ja aineistot
3. Arkistoaineiston erityispiirteet
4. Arkistoaineistolle räätälöidyn NER-mallin kehitystyö
5. Mitä seuraavaksi?



KANSALLISARKISTO

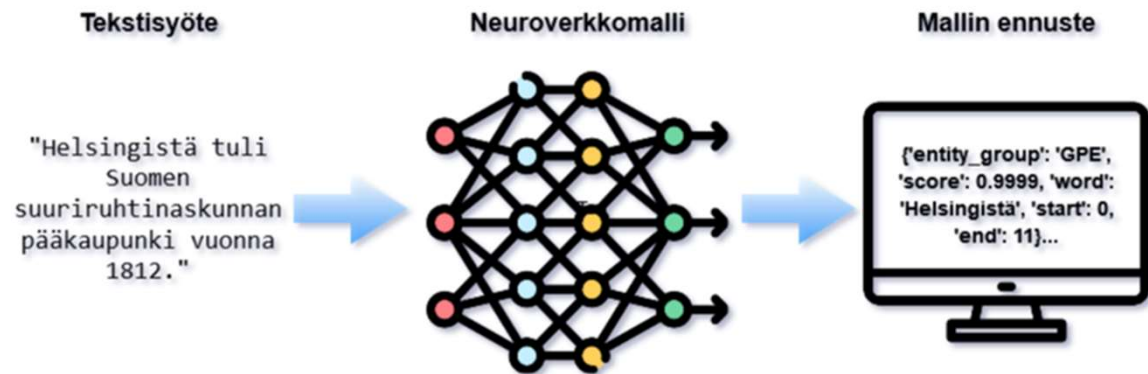
Kymmenenkään B-CARDINAL
 paikannussatelliittia 0
 ei 0
 vielä 0
 tarkoita 0
 , 0
 että 0
 Galileo B-PRODUCT
 olisi 0
 valmis 0
 . 0

 Lopullisen 0
 järjestelmän 0
 on 0
 määrä 0
 koostua 0
 24 B-CARDINAL
 operationaalisesta 0
 satelliitista 0
 ja 0
 kuudesta B-CARDINAL
 varasatelliitista 0
 vuonna B-DATE
 2020 I-DATE

Nimientiteettien tunnistus (NER): Mistä on kyse?

- Tunnistetaan tekstidatasta ennalta määriteltyihin kategorioihin kuuluvia elementtejä
- Yleisiä entiteettejä esim. henkilön nimet, paikannimet, organisaatiot, aikamääreet
- Helpottaa digitoidun aineiston automaattista metatiedottamista ja dokumenttien ja aineistojen linkittämistä, nopeuttaa tiedonhakua...

Nimientiteettien tunnistus (NER): Mistä on kyse?



Icons designed by Freepik and xnimrodx from Flaticon

- Nimientiteettien tunnistuksen kehitys: sääntöpohjaisista koneoppiviin malleihin, 2010-luvulta alkaen neuroverkkomallit hallinneet
- Nykyisin kielimallien hyödyntäminen yleistä: laajempi 'ymmärrys' kielestä auttaa tunnistamaan entiteettejä vaihtelevissa konteksteissa
- Englanti dominoinut, mutta myös suomenkielisiä malleja on kehitetty

Suomenkielinen NER: Mallit

- **FiNER** tagger: sääntöpohjainen NER-tunnistin
 - Lisätietoa:
<https://github.com/Traubert/FiNer-rules/blob/master/finer-readme.md>
 - Online-demo: <https://nlp.lda.fi/finer/>
 - Tunnistaa 22 entiteettiluokkaa (esim. 7 eri tyyppistä organisaatiota)

```
Valtion  
lisenssitoimikunnalle  
,  
  
Saksalainen  
Kokirjailija  
,  
tri      <EnamexPrsTit/>  
Friedrich      <EnamexPrsHum>  
Wolf      </EnamexPrsHum>  
)  
,  
Berlin-Pankow  
,  
on  
Suomen <EnamexLocPpl/>  
Yleisradiolle <EnamexOrgTvr/>  
luovuttanut  
2  
käsikirjoituksensa  
esitys-  
oikeuden
```

Osa FiNER taggerin tuloksista
esimerkkiasiakirjalle.

KANSALLISARKISTO

Suomenkielinen NER: Mallit

- TurkuNLP-tutkimusryhmän BERT-kielimallia hyödyntävä NER-tunnistin
 - Lisätietoa:
<https://turkunlp.org/fin-ner.html>
 - Online-demo:
<http://86.50.253.19:8001/tagdemo/>
 - Tunnistaa 18 entiteettiluokkaa

Legend
PERSON
NORP
GPE
ORG
QUANTITY

Valtion lisenssitoimikunnalle,
Saksalainen Kklrjailija, tri Friedrich Wolf), Berlin-Pankow,
on Suomen Yleisradiolle luovuttanut 2 käsikirjoituksensa esitysoikeuden, joista käsikirjoituksista toinen on jo esitetty ohjelmistossa ja toinen tullaan lähiaikoina esittämään. Esitysoikeuden luovuttamisesta on tri Wolf saamassa yhtiöltämme pakkionsa, jonka hän meille saapuneen tiedon mukaan toivoisi saavansa vastaanottaa elintarvikkeiden muodossa.
Tämän vuoksi Oy.Yleisradio Ab. anoo kunnioittavasti, että Valtion lisenssitoimikunta myöntäisi yhtiöllemme vientilisenssin 10 kgtlslle lihaa, sen lähettämistä varten tri Wolfille, Lisenssit pyydämme saada kahta 5 kgin lähetystä varten. |
Helsingissä, tammtikuun 20 pinä/1948. :
AA 7
Hella Wuolijoki. Einar Sundström.

TurkuNLP:n NER-mallin tuloksia
esimerkkiasiakirjalle.

KANSALLISARKISTO

Suomenkielinen NER: Aineistot

- **FiNER-korpus**

- <https://github.com/mpsilfve/finer-data>
- Digitoday-lehdestä poimittuja teknologia-alan artikkeleita (vuodelta 2014)
- 6 annotoitua entiteettiluokkaa

- **Turku NER-korpus**

- <https://github.com/TurkuNLP/turku-ner-corpus>
- mm. blogitekstejä, lakitekstejä ja uutisartikkeleita
- 6 annotoitua entiteettiluokkaa

EU:n	B-ORG	0	
viime	0	0	
vuonna	0	0	
valmistuneessa	0	0	
alustavassa	0	0	
selvityksessä	0	0	
Applen	B-ORG	0	
verojärjestelyt	0	0	
Irlannissa	B-LOC	0	
todettiin	0	0	
kestämättömäksi	0	0	
.	0	0	
Yhtiö	0	0	
on	0	0	
maksanut	0	0	
liikevaihdostaan		0	
Euroopassa	B-LOC	0	
alle	0	0	
kahden	0	0	
prosentin	0	0	
veroa	0	0	
kymmenen	0	0	
vuoden	0	0	

Esimerkki FiNER-aineistosta.

KANSALLISARKISTO

Suomenkielinen NER: Aineistot

- **Turku OntoNotes Entities Corpus**
 - <https://github.com/TurkuNLP/turku-one>
 - Sisältää aiemmat aineistot ja lisäksi eduskunnan laki- ja säädöstekstejä
 - 18 annotoitua entiteettiluokkaa
- **NewsEye-korpus**
 - <https://zenodo.org/record/4694466#.YJR20qE6-bi>
 - Sanomalehtiaineistoa 1850-1950-luvuilta
 - 4 annotoitua entiteettiluokkaa

Venäjän	B-LOC	O
suurempi	O	O
laivan	O	O
-	O	O
rakennusohjelma	O	O
.	O	O
V	O	O
.	O	O
1915	O	O
on	O	O
Englannin	B-LOC	O
dread	O	O
-	O	O
noughtien	O	O
lukumäärä	O	O
kohonnut	O	O
43:een	O	O
,	O	O
Saksan	B-LOC	O
25	O	O
-	O	O
27:ään	O	O
.	O	O

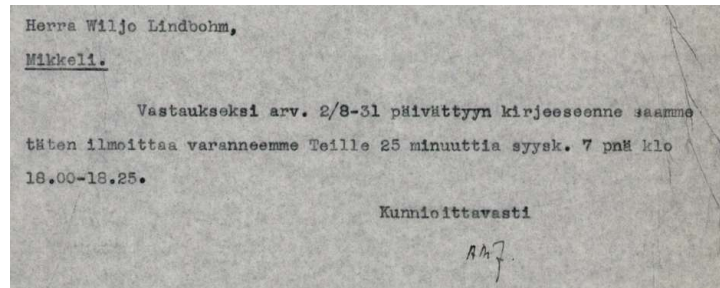
Esimerkki NewsEye-aineistosta.

KANSALLISARKISTO

Arkistoaineiston erityispiirteet

- Olemassa olevan aineiston ja NER-mallien rajoitteet
 - Mallien koulutuksessa ei ole käytetty juurikaan asiakirja-aineistoa
 - Koulutusdata on pitkälti digisyntyistä, kun taas digitoidun arkistoaineiston sisältö on tunnistettu optisen tekstintunnistuksen työkalujen avulla
 - Tekstintunnistus (OCR, HTR) tuottaa usein hälyä ja virheitä digitaaliseen tekstisisältöön: haaste malleille jotka on koulutettu 'hälyttömällä' datalla
 - Arkistoaineisto voi sisältää kiinnostavia entiteettikategorioita, joita olemassa olevia malleja ei kuitenkaan ole koulutettu tunnistamaan
- Motivaatio kehittää oma NER-malli erityisesti digitoitua asiakirja-aineistoa varten!

Arkistoaineiston erityispiirteet



Mikkeli.

,13 pni elokuuta = 31...

Herra Wiljo Lindbohm, N ;
N
j

Vastauksesi arv. 2/8-31 päivättyyn kirjeseenne SN S.
täten ilmoittaa varanneemme Teille 25 minuuttia syysk. 7 pñä klo a W :
18.00+18.25. | : | i 4 N

Kunnioittavasti Eeti W

An: dā s R

Mikkeli GPE ., 13 pni elokuuta = 31 DATE ... Herra Wiljo
Lindbohm PERSON , N ; N j Vastauksesi arv. 2/8-31 DATE päivättyyn
kirjeseenne SN S. täten ilmoittaa varanneemme Teille 25 minuuttia
syyske. 7 pñä DATE klo a W : 18.00+18.25. | : | i 4 N Kunnioittavasti
Eeti PERSON W An: dā s R Hella Wuolijoki PERSON . Einar
Sundström PERSON .

Digitoitu asiakirja.



OCR-sovelluksen (Tesseract) avulla
tunnistettu asiakirjan tekstisisältö.



Tunnistetun tekstisisällön pohjalta
löydetyt nimientiteetit.

KANSALLISARKISTO

Arkistoaineistolle räätälöidyn NER-mallin kehitystyö: Toimijat

- **DALAI** - Digitaalisten aineistojen laadun ja käytettävyyden parantaminen tekoälyavusteisesti
 - Projektin tavoitteena mm. kehittää digitoidun aineiston automaattista kuvailua tekoälyä hyödyntäen ja parantaa siten aineiston käytettävyyttä
 - Yhtenä osana nimientiteettien tunnistaminen asiakirja-aineistosta
- **FIN-CLARIAH** - valtakunnallinen ihmistieteiden tutkimusinfrastruktuuri
 - Yksi tavoitteista kehittää tutkijoiden avuksi tekoälypohjaisia ratkaisuja edistämään Kansallisarkiston digitalisoitujen tietoaineistojen tehokkaampaa käyttöä
- Yhteinen intressi nimientiteettien tunnistuksen kehittämiseen osana digitoidun asiakirja-aineiston käytettävyyden parantamista

KANSALLISARKISTO

Arkistoaineistolle räätälöidyn NER-mallin kehitystyö: Aineistot

- Projektissa annotoitu sekä uudempaa (mm. Työ- ja elinkeinoministeriön arkisto) että vanhempaa asiakirja-aineistoa (mm. Senaatin oikeusosaston päätöksiä vuodelta 1916)
- Lisäksi NewsEye-aineisto annotoitiin uudelleen, mukaan otettiin lisää entiteettikategorioita
- NER-mallin koulutuksessa käytetty myös Turku OntoNotes Entities Corpus-aineistoa
- Tavoitteena malli joka toimisi paitsi digitoidulla arkistoaineistolla myös yleisemmin eri tyyppisillä teksteillä

Keisarillinen	B-ORG
Senaatti	I-ORG
on	O
itselleen	O
esityttänyt	O
juttuun	O
kuuluvat	O
kirjat	O
ja	O
niistä	O
saanut	O
tietää	O
seuraavaa	O
:	O
Saarijärven	B-ORG
ja	I-ORG
Uuraisten	I-ORG
pitäjän	I-ORG
kärjäkunnan	I-ORG
Kihlakunnanoikeudessa	I-ORG
on	O

Esimerkki senaatin oikeusosaston aineistosta.

KANSALLISARKISTO

Arkistoaineistolle räätälöidyn NER-mallin kehitystyö: Aineistot

- Mallin koulutuksessa käytetyt aineistot:
 - Arkistoituja valtionhallinnon dokumentteja 1970-luvulta 2000-luvulle
 - Turku OntoNotes Entities ("TurkuONE") -korpus (aineisto 2000-luvulta)
 - NewsEye -korpuksen suomenkielinen osuus (sanomalehtiaineistoa 1850-1950-luvuilta)
 - Senaatin dokumentteja vuodelta 1916
 - Suomalaisia kirjoja Project Gutenbergista (1900-luvun alkupuolelta)
 - Suomalaisten ammattikorkeakoulujen lopputöitä 2000-luvulta

Arkistoaineistolle räätälöidyn NER-mallin kehitystyö: Entiteetit

- Annotoitavien entiteettien valinta tehtiin käyttäjien tarpeiden perusteella
- Kansallisarkisto lähetti kyselyn entiteeteistä 7 eri massadigitointiin osallistuvalla viranomaisella, Fin-Clariah tutkijat lähettivät saman kyselyn 58 eri tutkijalle
 - Vastaajat arvioivat eri entiteettien hyödyllisyyttä omasta näkökulmastaan
 - Vastausten perusteella annotoitaviksi valittiin 10 entiteettiä:

Arkistoaineistolle räätälöidyn NER-mallin kehitystyö: Entiteetit

- Henkilö (PERSON)
- Organisaatio (ORGANISATION)
- Paikka (LOCATION)
- Geopoliittinen alue (GEOPOLITICAL LOCATION)
- Tavara (PRODUCT)
- Tapahtuma (EVENT)
- Päivämäärä (DATE)
- **Diaarinumero** (JOURNAL NUMBER)
- **Y-tunnus** (FINNISH BUSINESS IDENTITY CODE)
- Kansallisuus, uskonnollinen tai poliittinen ryhmä (NATIONALITY, RELIGIOUS AND POLITICAL GROUPS)

Arkistoaineistolle räätälöidyn NER-mallin kehitystyö: Entiteetit

YY and YYYY represent years, AB and ABC represent any combination of letters.	YYYY and YY represent years.	AB can be written together or separately, both combinations of AB or ABC can be found.	DD=day, MM=month, YY and YYYY=year.
1/234/YY AB 123/456/YYYY AB 1234/567/YY AB 1234/56/YY AB 123/456/YYAB ABC/1234/01.01.01/YYYY ABC123/456/YYYY 123/456/YYYY 1234/567/YYYY	1/234/56 1/23/45 12/345/67 123/456/78 1234/567/89 12a/345/67 1/234/YYYY 1/23/YYYY 12/345/YYYY 123/456/YYYY 1234/567/YYYY 12a/345/YYYY 1-2/345/YY 1-2/34/ YY 12-34/567/ YY 123-456/789/ YY 1000-1100/123/ YY 1-2/345/YYYY 1-2/34/YYYY 12-34/567/YYYY 123-456/789/YYYY 1000-1100/123/YYYY	1/23/YYAB 1/23/YY AB 1/234/YYAB 1/234/YY AB 1/234/YYABC 1/234/YY ABC 12/345/YYAB 123/456/YYAB 1234/567/YYAB	ABC/123/45/YYYY ABC/1234/56/YYYY 12/34/56/DD.MM.YY AB 1234/ DD.MM.YY 12-345-67/DD.MM.YYYY 1234/567/89/6789 12/34/56/DD.MM.YY 123/45/67/DD.MM.YYYY AB 123:45/YYYY AB/12/34.56.78/YYYY 123/DD.MM.yy/YYYY ABCD-YYYY-01234/Ab-12 ABCDE-YYYY-01234/Ab-12cd ABCDE/123/04.05.06/YYYY AB-01-234-YYYY AB-01-234-YY

KANSALLISARKISTO

Arkistoaineistolle räätälöidyn NER-mallin kehitystyö: Annotointiprosessi

- FiNER-aineiston ja Turku NER-aineiston annotointiohjeiden pohjalta määriteltiin projektin omat annotointisäännöt
- Epäselvistä tapauksista käytiin keskustelua Discordissa
- Erityisesti valtionhallinnon (TEM) arkistoaineistossa haastavia annotoitavia olivat erilaiset organisaatioiden nimet ja tuote-entiteettiin kuuluvien tapausten täsmällinen määrittely

Arkistoaineistolle räätälöidyn NER-mallin kehitystyö: Annotointiprosessi

3.5 Product (PRODUCT)

Include:

- Artifacts: e.g. *iPhone 6*, *Lumia-puhelin*, *Hubble-avaruuskaukoputki* (*Hubble telescope*), *Compute Stick*, *Playstation*, *Snapdragon-suoritin*, *Watsonsupertietokone*
- Laws/initiatives: e.g. *Patriot Act -laki* (*the Patriot Act law*), *Järkeä tekijänoikeuslakiin - kansalaisaloite* (*Make sense to copyright law citizen initiative*)
- Networks (other than television network): e.g. *Tor*
- Platforms: e.g. *Google Play*, *Kickstarter*
- Product series/collections: e.g. *-sarja*, *-mallisto*
- Programming languages: e.g. *Java*, *C++*
- Projects/Programs/Operations: e.g. *Vitja*, *VideoLan-projekti*, *Blue Shield - sairausvakuutusohjelma*, *Operation Pawn Storm*
- Protocols: e.g. *pop3*, *imap*, *VPN*
- Published copies and issues: e.g. *Helsingin Sanomat*
- Services/Platforms: e.g. *Apple Store*, *Google Play*, *Tor Mail* -sähköpostipalvelu
- Slogans: e.g. *"Don't be evil"*
- Software: e.g. *Windows 10*, *Trojan-Banker.Win32.Chthonic*, *Dropbox*
- Systems: e.g. *Alipay-järjestelmä*
- Technologies: e.g. *HoloLens-teknologia*
- Vehicles/Vessels: e.g. *Tesla Model S*, *Opportunity*, *Kansainvälinen avaruusasema* (*International Space Station*), *Helsingin Metro*
- Websites, forum: e.g. *Amazon.com-sivusto*, *Tvkaista.fi-verkko-osoite*, *Suomi24*, *Futisfoorumi*, *Ylilauta*
- Works/Art: e.g. *Tuntematon sotilas*, *Hurt Locker*, *Angry Birds*, *American Idol*
- Named agreements, treaties and laws are annotated as PRO: e.g. *Euroopan yhteisön perustamissopimus*, *Kioto sopimus*, *Suomen perustuslaki*, *GDPR*, *KVTES*

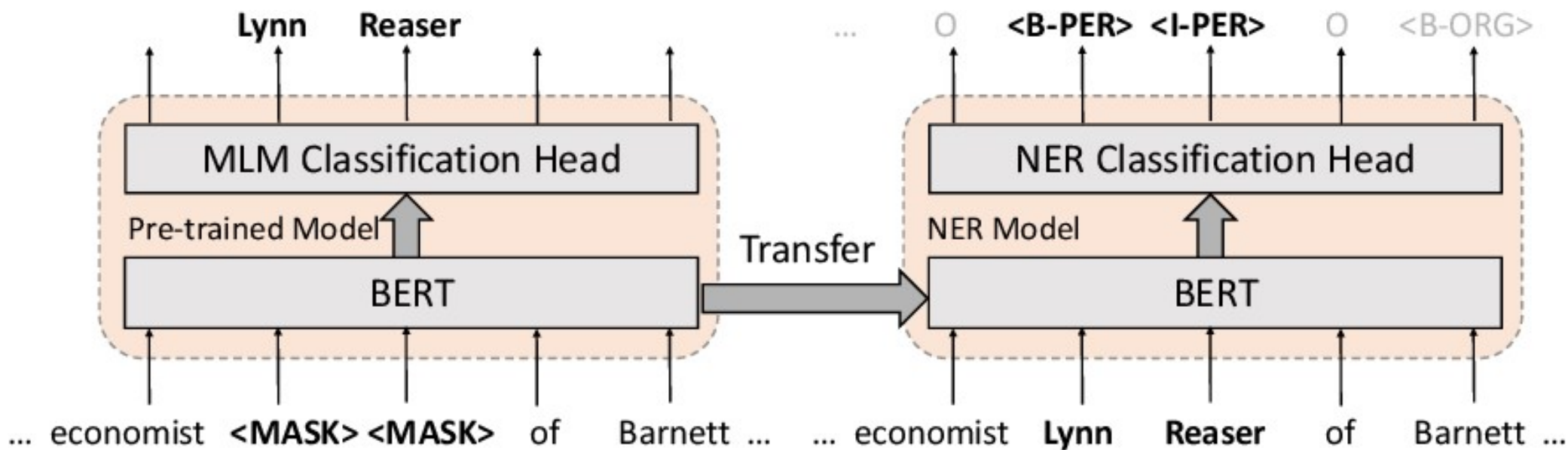
Esimerkki annotointiohjeista: PRODUCT-kategoriaan kuuluvia tuotetyyppejä.

KANSALLISARKISTO

Arkistoaineistolle räätälöidyn NER-mallin kehitystyö: Mallin koulutus

- Mallin kouluttamisessa hyödynnettiin FinBERT-kielimallia sekä HuggingFace-kirjastoa
- Mukana monipuolisesti annotoitua koulutusaineistoa
- Yhteensä yli 128 000 annotoitua entiteettiä
- Mallin koulutuksessa käytettiin Kansallisarkiston 2 GPU:lla varustettua 'tehotyöasemaa': ei vaadi supertietokonekokoluokan resursseja!

Arkistoaineistolle räätälöidyn NER-mallin kehitystyö: Mallin koulutus



Kuva artikkelista Liang, C., Yu, Y., Jiang, H., Er, S., Wang, R., Zhao, T., & Zhang, C. (2020). BOND: BERT-Assisted Open-Domain Named Entity Recognition with Distant Supervision. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

KANSALLISARKISTO

Arkistoaineistolle räätälöidyn NER-mallin kehitystyö: Tuloksia

- Tulokset vaihtelevat eri entiteettiluokkien osalta: esim. henkilönnimet, geopoliittiset paikannimet ja päivämäärät tunnistuvat paremmin kuin tapahtumat ja tuotteet
 - Heijastelee ainakin osittain näiden entiteettien määriä koulutusaineistossa

	Precision	Recall	F1-score
PERSON	0.90	0.91	0.90
ORG	0.84	0.87	0.86
LOC	0.84	0.86	0.85
GPE	0.91	0.91	0.91
PRODUCT	0.73	0.77	0.75
EVENT	0.69	0.73	0.71
DATE	0.90	0.92	0.91
JON	0.83	0.95	0.89
FIBC	0.95	0.99	0.97
NORP	0.91	0.95	0.93

Mallin tuloksia kuvaavia metriikoita.

KANSALLISARKISTO

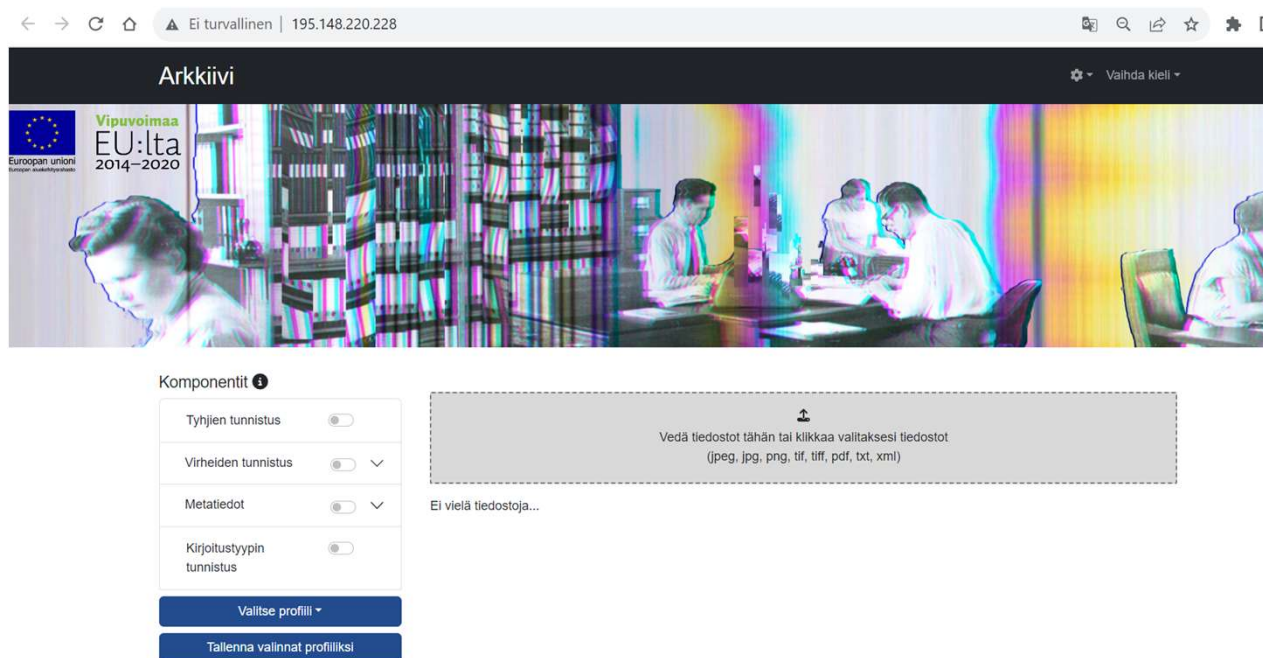
Arkistoaineistolle räätälöidyn NER-mallin kehitystyö: Tuloksia

- Malli joka on koulutettu sekä uudella digisyntyisellä aineistolla että OCR:ätyllä arkistoaineistolla vaikuttaisi tunnistavan entiteettejä digisyntyisestä aineistosta vastaavalla tarkkuudella kuin pelkästään sillä koulutettu malli
- Selkeä ero tunnistuksen laadussa digitoidun OCR:ätyn arkistoaineiston osalta: tulokset huomattavasti parempia silloin kun koulutusaineisto sisältää myös vastaavan tyyppistä arkistoaineistoa

Arkistoaineistolle räätälöidyn NER-mallin kehitystyö: Käyttö

- Malli on saatavilla Kansallisarkiston HuggingFace-sivuilla, jossa sitä voi myös testata: <https://huggingface.co/Kansallisarkisto/finbert-ner>
- Malli on myös mukana Arkkiivi-palvelussa: <https://arkkiivi.fi/>
- Lisäksi GitHubissa on saatavilla
 - API-versio mallista: https://github.com/DALAI-project/NER_API
 - Mallin koulutuksessa käytetty koodi: https://github.com/DALAI-project/Train_BERT_NER

Arkistoaineistolle räätälöidyn NER-mallin kehitystyö: Käyttö



Arkkiivi-käyttöliittymä sisältää komponentteja jotka tunnistavat digitoiduista aineistoista skannausvirheitä ja sisältöjä.

KANSALLISARKISTO

Mitä seuraavaksi?

- Nykyisessä mallissa mukana vain vähän vanhempaa aineistoa, suunnitelmissa laajentaa aineistoa ajallisesti ja/tai tuottaa oma NER-malli vanhemmalle arkistoaineistolle
- Entiteettien linkittäminen esim. geotietoon tai muuhun aineiston ulkopuoliseen informaatioon tarjoaa myös monenlaisia mahdollisuuksia



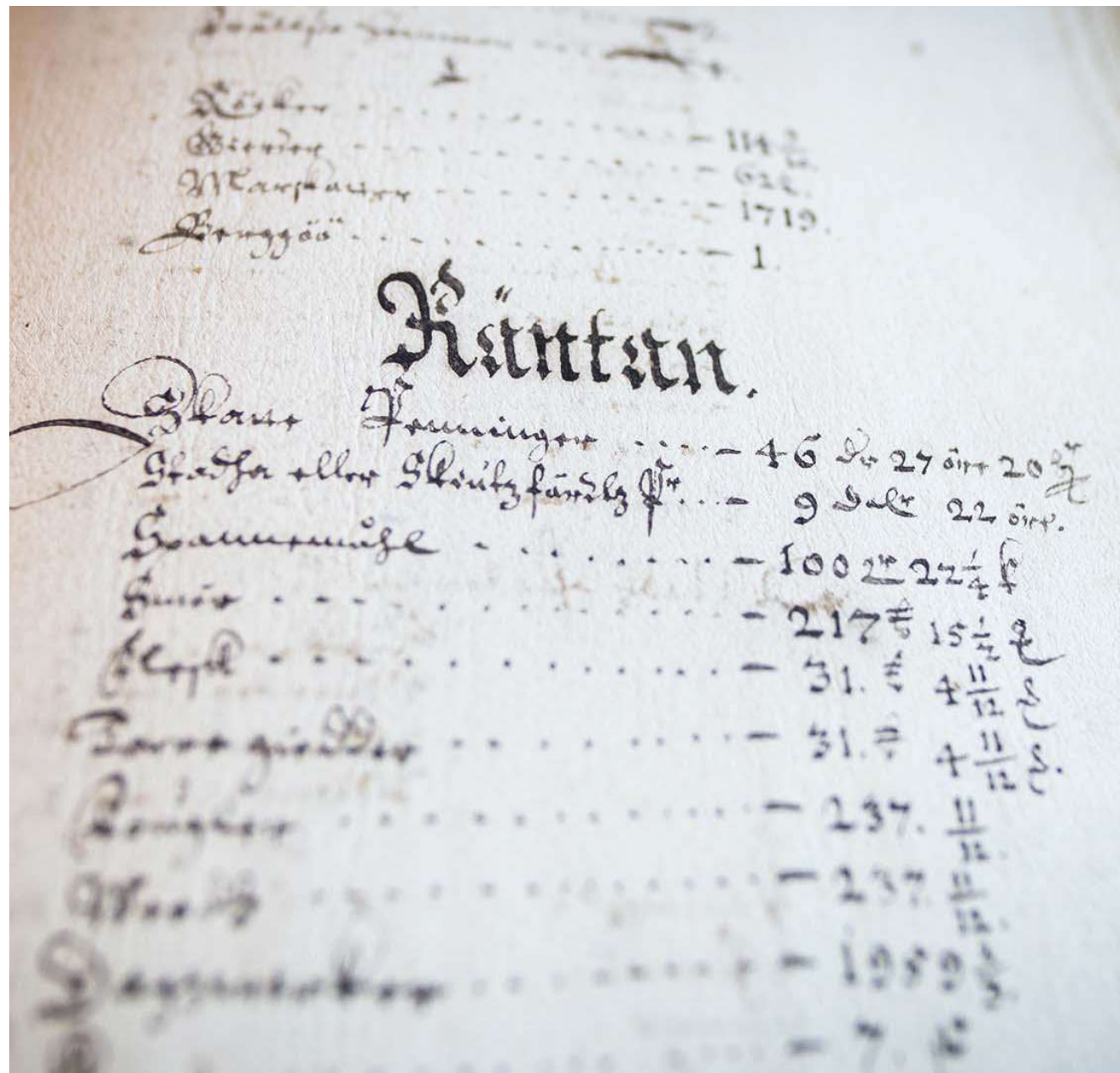
KANSALLISARKISTO



KANSALLISARKISTO

Kiitos!

mikko.lipsanen@kansallisarkisto.fi





KANSALLISARKISTO

www.kansallisarkisto.fi



@kansallisarkisto



@kansallisarkist