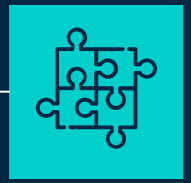
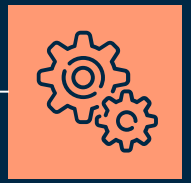
Nimientiteettien tunnistusmallin kehitystyö

Mikko Lipsanen
Koneoppimisen pääsuunnittelija
Kansallisarkisto



01

Tausta ja motivaatio



02

Aineisto ja annotointi

03

Mallin kouluttaminen ja tulokset

Tausta ja motivaatio

01

Nimientiteettien tunnistus (NER)

- Mistä NERissä on kyse?
  - Tunnistetaan tekstidatasta ennalta määriteltyihin kategorioihin kuuluvia elementtejä
  - Yleisiä entiteettejä esim. henkilön nimet, paikannimet, organisaatiot, aikamääreet..
  - Hyödyllinen työkalu digitoidun aineiston automaattisessa metatiedottamisessa

NER-kehitystyö: toimijat

- **DALAI** - Digitaalisten aineistojen laadun ja käytettävyyden parantaminen tekoälyavusteisesti
  - 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

NER-kehitystyö: toimijat

- Yhteinen intressi nimientiteettien tunnistuksen kehittämiseen osana digitoidun asiakirja-aineiston käytettävyyden parantamista
- Kehitystyötä tehty DALAI-hankkeen ja FIN-CLARIAHin /Jyväskylän yliopiston humanistis-yhteiskuntatieteellisen tiedekunnan yhteistyönä

Aineisto ja annotointityö

02

Aineistot

- Suomenkieliset NER-aineistot
  - **Finer-korpus**
    - https://github.com/mpsilfve/finer-data
    - Koostuu Digitoday-lehdestä poimituista teknologia-alan artikkeleista (vuodelta 2014)
    - 953 artikkelia, 193 742 tokenia
    - 6 annotoitua entiteettiluokkaa (ORG, LOC, PER, PROD, EVENT, DATE)
  - **Turku NER-korpus**
    - https://github.com/TurkuNLP/turku-ner-corpus
    - Sisältää mm. blogitekstejä, lakitekstejä ja uutisartikkeleita
    - 754 dokumenttia, 200 000 tokenia
    - 6 annotoitua entiteettiluokkaa

Aineistot

- **Turku OntoNotes Entities Corpus**
  - https://github.com/TurkuNLP/turku-one
  - N. 500 000 tokenia
  - Sisältää FiNER- ja Turku NER-aineiston sekä 60 eduskunnan lakia ja säädöstä
  - 18 annotoitua entiteettiluokkaa
- **NewsEye-korpus**
  - https://zenodo.org/record/4694466#.YJR20qE6-bi
  - N. 60 000 tokenia
  - OCR:ättyä sanomalehtiaineistoa 1850-1950 -luvuilta
  - 4 annotoitua entiteettiluokkaa

```
Pelkkää O
tyhjyyttä       O

Kävin   O
tänään  B-DATE
katsomassa      O
Suomen  B-ORG
Perinteisen     I-ORG
Teatterin       I-ORG
näytelmän       O
Ranta   B-WORK_OF_ART
.       O

Jo      O
teatterin       O
nimi    O
antikliimaksi   O
;       O
ikäänkuin       O
ryhmä   O
tahtoisi        O
antaa   O
katsojalle      O
vakuutuksen     O
siitä   O
```

Esimerkki Turku OntoNotes Entities Corpus-aineistosta

Aineistot

- Olemassa olevan aineiston rajoitteet
  - Ei sisällä juurikaan asiakirja-aineistoa
  - Ei sisällä juurikaan OCR:ättyä aineistoa
- Projektissa annotoitu sekä uudempaa OCR:ättyä asiakirja-aineistoa (mm. Työ- ja elinkeinoministeriön arkisto) että vanhempaa HTR:ättyä aineistoa (mm. Senaatin oikeusosaston päätöksiä vuodelta 1916)
- Lisäksi NewsEye-aineistoa annotoitiin uudelleen, mukaan otettiin lisää entiteettikategorioita
- NER-mallin koulutuksessa käytetty myös Turku OntoNotes Entities Corpus-aineistoa

Entiteetit

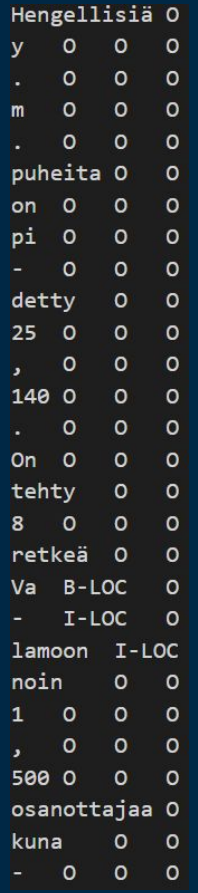
- Kansallisarkisto lähetti kyselyn entiteeteistä 7 eri massadigitointiin osallistuvalle viranomaiselle, Fin-Clariah tutkijat lähettivät saman kyselyn 58 eri tutkijalle
    - Vastaajat arvioivat eri entiteettien hyödyllisyyttä omasta näkökulmastaan
    - Vastausten perusteella annotoitaviksi valittiin 10 entiteettiä:

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)

Annotointiprosessi

- FiNER-aineiston ja Turku NER-aineiston annotointiohjeiden pohjalta määriteltiin projektin omat annotointisäännöt
- Epäselvistä tapauksista käytiin keskustelua Discordissa
- Annotointi jatkuu elokuun loppuun, tähän mennessä itse annotoidussa aineistossa jo yli 1 200 000 tokenia
    - Huom! OCR pilkkoo usein sanoja pieniin osiin, joten token-määrä ei ole verrannollinen aiemmin käsiteltyihin aineistoihin (joissa se vastaa paremmin aineiston sanamäärää)



```
Hengellisiä O
y   O   O   O
.   O   O   O
m   O   O   O
.   O   O   O
puheita O   O
on  O   O   O
pi  O   O   O
-   O   O   O
detty   O   O
25  O   O   O
,   O   O   O
140 O   O   O
.   O   O   O
On  O   O   O
tehty   O   O
8   O   O   O
retkeä  O   O
Va  B-LOC   O
-   I-LOC   O
lamoon  I-LOC
noin    O   O
1   O   O   O
,   O   O   O
500 O   O   O
osanottajaa O
kuna    O   O
-   O   O   O
```

Esimerkki NewsEye-aineistosta

Mallin kouluttaminen ja tulokset

03

NER-tunnistuksen työkalut

- Mitä työkaluja olemassa suomen kielellä?
  - **FiNER** tagger: sääntöpohjainen työkalu
    - Lisätietoa: https://github.com/Traubert/FiNer-rules/blob/master/finer-readme.md
    - Online-demo: https://nlp.ldf.fi/finer/
    - Tunnistaa 22 entiteettiluokkaa (esim. 7 eri tyyppistä organisaatiota)
  - **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

BERT-NER

- Kehitys sääntöpohjaisista kopeoppiviin malleihin, viime vuosina kielimallien hyödyntäminen parantanut tuloksia
- Kielimallit koulutetaan suurilla aineistoilla oppimaan kielen rakenteita kuvaavia todennäköisyyksiä: näitä pohjamalleja voidaan jatkokouluttaa erilaisiin spesifeihin tehtäviin, kuten nimientiteettien tunnistus
- **FinBERT:** TurkuNLP-ryhmän kouluttama suomenkielinen BERT-malli
  - Koulutusaineistossa yli 3 miljardia tokenia
  - Ryhmä saavutti BERT-pohjaisella mallilla selkeästi verrokkimalleja parempia tuloksia nimientiteettien tunnistuksessa*

*Jouni Luoma, Miika Oinonen, Maria Pyykönen, Veronika Laippala, Sampo Pyysalo. 2020. A Broad-coverage Corpus for Finnish Named Entity Recognition. In Proceedings of The 12th Language Resources and Evaluation Conference (LREC'2020).)

BERT-NER

- Hyödynsimme myös FinBERT-mallia sekä HuggingFace-kirjastoa oman NER-mallin kouluttamisessa
- Aineiston annotointi vielä kesken, valmistuu syyskuun alussa jonka jälkeen koulutetaan uusi malli
- Viimeisin malli 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

Mallin koulutus ja testaus

- Koulutuksessa hyödynnetty Kansallisarkiston GPU-resurssia
- Tulokset vaihtelevat eri entiteettiluokkien osalta: esim. henkilönnimet, geopoliittiset paikannimet ja päivämäärät tunnistuvat paremmin kuin tapahtumat, y-tunnukset ja diaarinumerot
  - Heijastelee ainakin osittain näiden entiteettien määriä koulutusaineistossa
- Eri aineistotyyppien (OCR:ätty vs. manuaalisesti transkriboitu teksti, uusi vs. vanha aineisto) tunnistuksen eroja tutkittu vasta alustavasti
  - Parhaiten vaikuttaisi tunnistuvan uusi manuaalisesti transkriboitu aineisto, huonoiten vanha OCR:ätty aineisto: tähän vaikuttanee osaltaan BERT-pohjamallin koulutuksessa käytetty data

Mallin jatkokehitys

- Syyskuussa koulutettava uusi malli sisältää aiempaa enemmän sekä OCR:ättyä asiakirja-aineistoa että myös HTR:ättyä 1900-luvun alun senaatin aineistoa
- Tutkitaan miten hyvin malli toimii historiallisen aineiston kanssa, mahdollista tuottaa jatkossa lisää koulutusmateriaalia hyödyntäen vanhempaa asiakirja-aineistoa
- Aiheeseen liittyen ollaan myös työstämässä syksyn aikana tutkimusartikkelia

Kiitos!

mikko.lipsanen@kansallisarkisto.fi