Aalto University

Automaattinen semanttinen annotointi

Matias Frosterus, Reetta Sinkkilä, Katariina Nyberg
Semantic Computing Research Group (SeCo)
Aalto University School of Science and Technology,
Department of Media Technology
and
University of Helsinki, Department of Computer Science

Semanttinen annotointi

Aalto University

- Aineiston kuvailua ontologisilla käsitteillä
  - Liitetään eksplisiittisesti kohteeseen jonkin ontologian käsitteitä
- Mahdollistaa päättelyn ja korkeamman tason automaation kohteita käsiteltäessä
  - koneluettava
  - voidaan esim. määrittää että dokumentissa esiintyvä merkkijono ”Pariisi” merkitsee yksiselitteisesti paikkaontologiassa määriteltyä käsitettä, joka vastaa Teksasissa sijaitsevaa kaupunkia
- Voidaan tehdä käsin
  - esimerkiksi SAHA-työkalulla

Motivaatio

Aalto University

- Käsin annotointi on hidasta
- Automaatio on välttämätöntä kun
  - halutaan annotoida suuria olemassa olevia kokoelmia
  - ei haluta annotoida käsin
    - » ”ei kuulu työnkuvaan”

Automaattinen semanttinen annotointi (1/2)

Aalto University

- Etsitään vastaavuuksia kohteessa esiintyvien ja ontologian käsitteiden nimien merkkijonojen välillä
  - lemmataan (saatetaan sanat perusmuotoon) annotoitava teksti
  - lemmataan ontologian käsitteiden nimet
  - etsitään vastaavuudet

Automaattinen semanttinen annotointi (2/2)

Aalto University

- Annotointiskeema määrittelee mitä metatietoja kohteesta halutaan
  - tietyn ominaisuuden arvojoukkoa voidaan rajoittaa
    - » esim. valmistusmaa on aina valtiot-luokan ilmentymä
  - kohteen olemassa olevaa rakenteellisuutta, jos sellaista on, voidaan myös hyödyntää annotoinnissa
    - » esim. XML-dokumentin <author>-elementtien välistä löydetyt käsitteet ovat dokumentin tekijöitä

Automaattisen semanttisen annotoinnin haasteita (1/3)

Aalto University

- Disambiguointi
  - ontologisten käsitteiden välillä vs. käsitteen ja ei-käsitteen välillä
  - ontologiassa maailma hahmotetaan käsitteillä, eikä niitä tarkoittavilla sanoilla
    - » yksi termi voi vastata montaa käsitettä
    - » esim. lapset ikäryhmänä tai sukulaisuussuhteena
  - miten erotetaan mistä käsitteestä on kyse?
    - » kontekstitiedolla
      - etukäteissäädöillä (”tässä aineistossa lapset viittaa aina ikäryhmään”)
      - päättelemällä ympäröivistä käsitteistä
      - koneoppimisella tms.

Automaattisen semanttisen annotoinnin haasteita (2/3)

Aalto University

- Väärät annotaatiot
  - esim. ”tämä artikkeli ei käsittele arkeologiaa” voisi tuottaa annotaatioksi arkeologia-käsitteen
    - » voidaan koettaa ratkaista syntaktisella jäsennyksellä

Automaattisen semanttisen annotoinnin haasteita (3/3)

Aalto University

- Uusien instanssien löytäminen
  - esim. uusien henkilöiden löytäminen
    - » voidaan etsiä etunimi-sukunimipareja
      - tunnetut etunimet (etunimiluettelo)
      - isot alkukirjaimet
      - morfologinen päätevaihtelu (etunimi ei taivu kun se esiintyy lauseessa ennen sukunimeä: ”Matti Järvisellä oli”)
    - » löydetyn parin jälkeiset viittaukset samaan yksittäiseen etunimeen tai sukunimeen tulkitaan viittauksiksi löydettyyn instanssiin

Poka

Aalto University

- Poka on automaattisen annotoinnin perustyökalu ontologisten käsitteiden (RDF, OWL, SKOS) etsimiseen tekstistä
  - käyttää Connexorin kielityökalua
  - sisältää myös käsiteriippumattoman henkilöeristimen ja säännöllisten lausekkeiden eristimen

Airo

Aalto University

- Kohteena Helsingin Sanomien sähköisessä muodossa oleva artikkeliarkisto
  - tehokkaampi haku
- Automaattinen annotointi Pokalla YSOa vasten
- Dokumentin laajennos
  - dokumenttien asiasanoitukseen liitettiin sellaisia käsitteitä, joilla oli ontologiassa läheiset suhteet automaattisessa annotoinnissa löydettyihin käsitteisiin
  - tulokset positiivisia

Vanhat sanomalehtiaineistot

Aalto University

- Kansalliskirjaston historiallinen sanomalehtiarkisto
  - kaikki Suomessa ilmestyneet sanomalehdet vuosilta 1771-1890
  - 400 000 otsikkoa
- Automaattinen semanttinen annotointi poikkeavuuksia sisältävälle aineistolle
  - kirjoitusasut, vanhat sanat, OCR-virheet, lyhenteet, ...
  - pyritään arvaamaan oikea muoto sanalle etsimällä melkein samanlaisia merkkijonoja ontologiasta ja sanakirjasta
    - » valinta useiden samankaltaisten sanojen joukosta käyttäen sääntöjä, jotka perustuvat tunnettuihin virheisiin
      - Wenäjä – Venäjä
      - lutherilainen – luterilainen
      - kapteini – kapteeni

UNIVERSITY OF HELSINKI

SeCo semantic computing

SÄHKE-asiakirjojen automaattinen luokittelu tekstin pohjalta

Aalto University

- SÄHKE on Kansallisarkiston sähköisen arkistoinnin metadatamalli
- SÄHKE-asiakirjat annotoitiin Pokalla
  - ontologiana YSO
- Automaattinen luokittelu oppivan mallin avulla
  - Annotaatioiden käyttö paransi luokittelua

Blogilouhinta

Aalto University

- Tarkoitus: suositellaan Web 2.0-hengessä TerveSuomi-portaalin sisältöihin liittyviä tekstejä
  - blogit
  - uutissyötteet
  - keskustelupalstat
- Annotoidaan Pokalla ja tuotetaan TerveSuomen metatietomäärittelyn mukaista RDF:ää
  - asiasanoitus (dc:subject) Terveyden edistämisen ontologian käsitteillä

Kiitos

Aalto University

- Kysymyksiä?

UNIVERSITY OF HELSINKI

SeCo
semantic computing