Semantic Webin ontologiat ja sanojen disambiguaatio

Harri M. T. Saarikoski

Helsingin yliopisto / AAC Global

@ Älyä verkossa -symposio

Disambiguaatio-ongelma

- Disambiguaatiomenetelmillä pyritään automaattisesti ratkaisemaan tekstin **monimerkityksiset sanat yksiselitteisiin käsitteisiin**.
  - esim. ”channel” – tv-kanava vai Englannin kanaali
  - esim. ”Liverpool” – jalkapallojoukkue vai kaupunki (erisnimet => Named Entity Recognition)

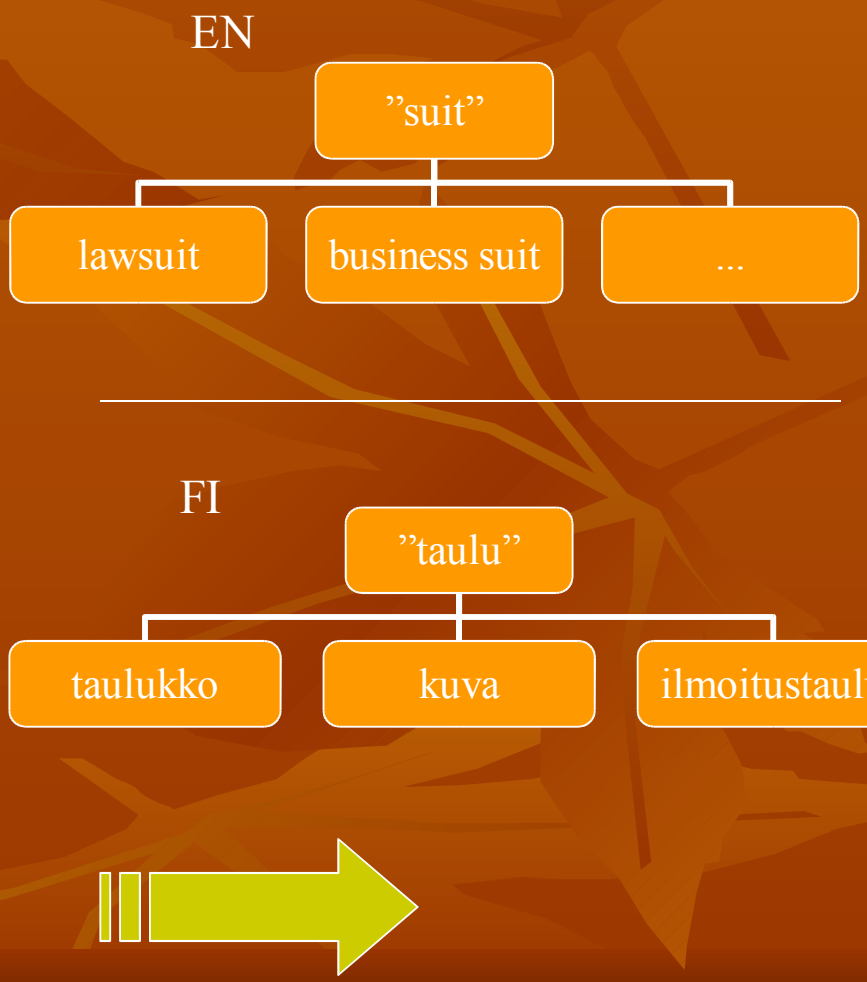
- Ongelman laajuus:
  - **1/3 englannin sanoista** on monimerkityksisiä.

- Luotettavien **tiedonirrotus-, tekstinymmärrys-, konekäännös- ja tiedonhakusovellusten** toteuttaminen vaatii ’toimivan’ disambiguaatiokomponentin.
  - esim. tiedonhaussa hakutermien disambiguoinnilla on keskimäärin **7-14%** hakusovelluksen tarkkuutta parantava vaikutus (Schütze ja Pedersen 1995).

Onko sellaista ?

Disambiguaatiojärjestelmät

- Tarkkuus on **65-75%:n** paikkeilla.
  - Aikaansaatu yleispätevällä, hienojakoisella merkitysjaolla.
  - Moniin sovelluksiin riittää karkeampi jako.
- Parhaat tulokset v. 2001 saatiin ’roolittamalla’
  - **yhteisesiintymälaskureita**
  - **monta eri luokittelijaa**
  - **eri käsitetietämyslähteitä**
    - *Olemassaoleva* käsitetietämys ei aina auttanut, paikoin jopa johti harhaan.



Disambiguaatioresursseja relaatiotyypeittäin

- Paljon tutkitut eli siis ei-riittävät:
  - car#1 occursOftenWith driver,accident...(automatic)
  - car#1 isTypeOf vehicle
  - car#1 hasType Sedan
  - car#1 hasPart wheel#3
  - car#1 hasUser(type:human)
  - car#1 hasFeature isFast
  - car#1 partOfDomain motorsports...
- Vähän tutkitut tai ei-olemassaolevat:
  - car#1 partOfEvent traffic,buyingCar,wreckingCar...
  - car#1 hasFunction driving,transportingHumans

SW-ontologiat ?!
SW-ontologiat ?!

SW-ontologioiden sisältö – ennuste

- Suunnitellut SW-sovellukset mahdollistaakseen ontologioissa on oltava
  - **tiheä käsiteobjektiverkko liitettynä yleisontologiaan**
  - **monenlaiset ja kattavat relaatiotyypit**
  - **päättelysääntöjä, 'maailman tietämystä'**

"...semantic indexing,
data integration,
knowledge formation,
content highlighting,
information retrievability..." (
www.semanticweb.org sovellusselvitys)

SW-ontologioiden käyttö disambiguaatiossa – 1

- **Disambiguoidaan ontologisoidun webbisivun teksti**
  - Koska ontologia on valmiiksi disambiguoitu/yksiselitteinen resurssi, itse ontologisoidun webbisivun teksti voidaan automaattisesti disambiguoida.
  - Tästä saadaan järjestelmille ’melko’ varmaa merkitystägättyä harjoitusdataa muidenkin tekstien disambiguointiin.

Bootstrapping !

SW-ontologioiden käyttö disambiguaatiossa – 2

- **Disambiguoidaan tekstejä** ontologioiden avulla
  - Ontologia on käsitteiltään ja relaatioiltaan yksiselitteinen.
  - Ei ole tutkimustuloksia.
    - Erityisesti niiltä osin disambiguaatiovoima on tutkittava kun käytetään aiemmin disambiguaatiossa tutkimattomia relaatiotyyppejä.

Lisäpiirteinä disambiguaatio-menetelmälle...

SW-ontologioiden käyttö disambiguaatiossa – 3

- **Hankitaan merkitystägättyä korpusta**
  - Mitä enemmän harjoitusdataa, sen paremmaksi järjestelmät tulevat.
  - Mihalcea ja Moldovan (2001) onnistuivat bootstrapping-menetelmällä irrottamaan internetistä merkitystägättyä korpusta valikoiduille sanoille
    - Käyttivät yksimerkityksisiä sanoja (WordNet) 'ankkurikohtina'.
    - WordNetin käsitemääritelmät annettiin pohjapiirteinä.
    - Saatiin **riittävä (92%) disambiguaatiotarkkuus** sanaesiintymille internetissä.

Tehtävissä siis !

SW-ontologioiden käyttö disambiguaatiossa – 4

- **Ontologian irrottaminen tekstistä**
  - Vaatii monen yhteenkietoutuneen tason disambiguaatiota:
    - sanat käsitteiksi ja
    - lauseet objektien välisiksi relaatioiksi
    - löytää todenmukaiset relaatiot irrotettujen joukosta
  - Omelayenko 2001 on julkaissut ontologioiden oppimismetodien katselmuksen.

Hankalampaa...

SW-ontologioiden käyttö disambiguaatiossa - 5

- **Etsitään internetistä rinnakkaiskorpuksia**
  - Koska ontologia on disambiguoitu resurssi, sen voi **monikielisen yleisontologian** (WN) avulla kääntää toiselle kielelle.
  - Näiden monikielisten ontologioiden avulla voidaan verkosta löytää (osittaisia) **rinnakkaiskorpuksia.**

Yhteenveto

- Tekstin automaattista yksiselitteistystä haittaa kumulatiivisesti **disambiguaatiovälitehtävien epätarkkuus**
  - sanojen ja rakenteiden monimerkityksisyys
  - moniosaisten käsitteiden tunnistus

- Ontologioiden vaikutusta ei ole voitu tutkia, vain ennustaa.
  - Ontologiat kuitenkin **laajentavat aihealuekattavuutta ja tuovat käytettäväksi monia käsitetietämysresurssien relaatioita laajassa mitassa**.
  - Jää nähtäväksi kuinka hyödyllisiä SW-ontologiat ovat edellämainittuihin disambiguaatiotehtäviin.

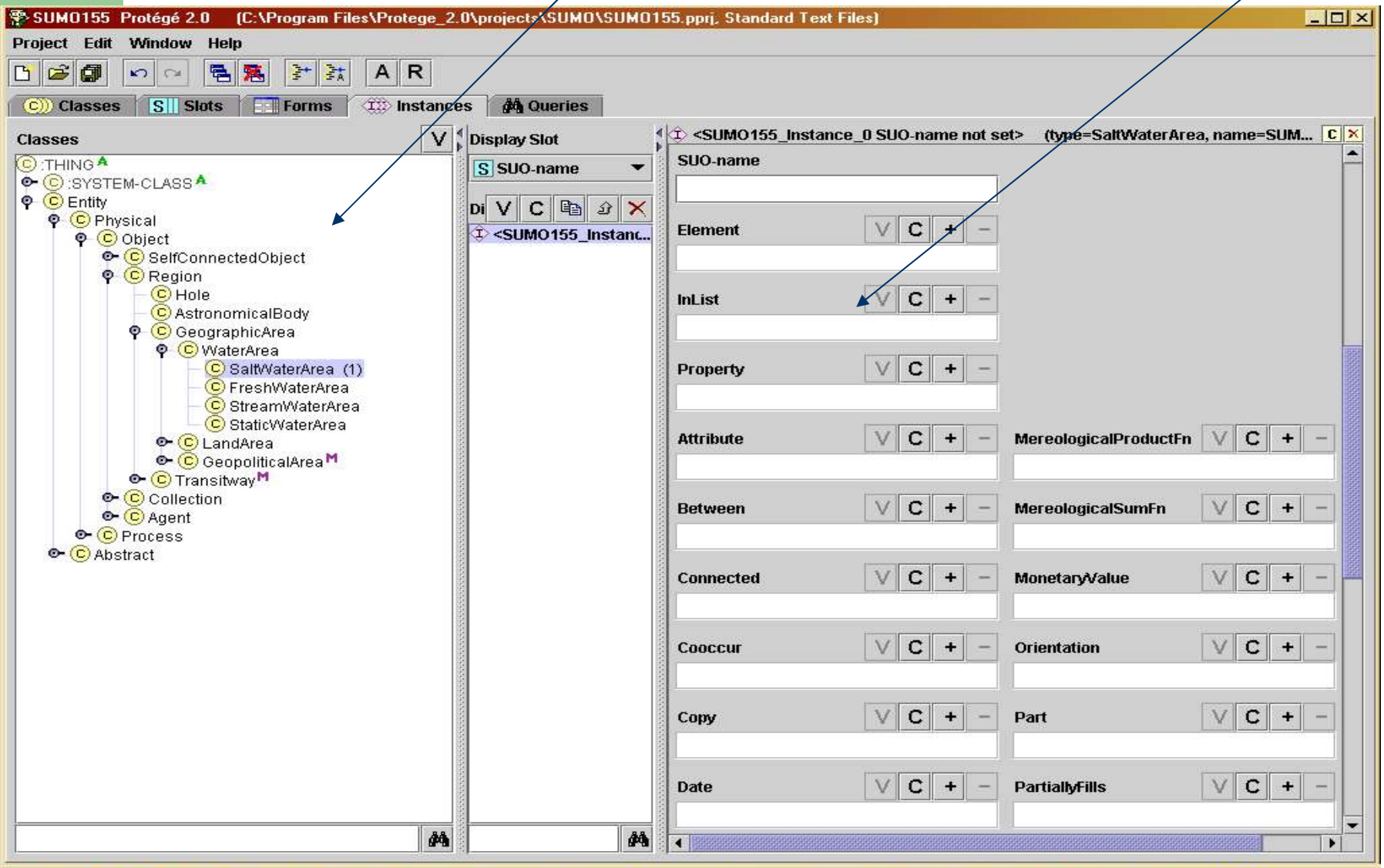
Lähteitä ja aktiviteetteja

- **Resurssit**
  - **Semantic Web, SW www.semanticweb.org**
    - Ontologioiden yhteensopivuus: http://www-db.stanford.edu/SKC/
    - Sovellusselvitys: http://www.w3.org/2001/sw/Europe/reports/chosen_demos_rationale_report/hp-applications-survey.html
  - **ebXML** (enabling Business with XML): joidenkin alojen ydinkäsitteet on koodattu standardiksi interoperabilityn aikaansaamiseksi http://www.ebxml.org/
  - **Google**: sivujensa mainostajat saavat nykyisin halutessaan tarkentavan hakusarakkeen (koeta esim. hakua ”ebXML”, tarkentava haku ilmestyy oikealle) www.google.fi
  - **SUMO (Suggested Upper-level Merged Ontology)** – laaja yleisontologia ehdolla Semantic Webin domain-ontologioiden pohjaksi http://ontology.teknowledge.com/
  - **WordNet** – laaja disambiguaatiomenetelmien kehityksessä käytetty käsiteverkko http://www.cogsci.princeton.edu/cgi-bin/webwn
  - **SENSEVAL-sivusto www.senseval.org** sisältää disambiguaatiokorpuksiin ja –resursseihin sekä järjestelmien evaluaatiotulokset ja kuvaukset.
  - **Open Mind Word Expert**: hajautettu merkitystägättyjen korpusten keräysprojekti SW-ontologioiden tyyliin http://www.teach-computers.org/word-expert.html
  - **MikroKosmos-ontologia:** huomattava määrä kovakoodattua käsitetietämystä mahdollisti yli 95%:n tarkkuuden espanja-englanti-konekäännöksissä http://ilit.umbc.edu/
- **Viitteet**
  - Mihalcea ja Moldovan 2001. A Highly Accurate **Bootstrapping** Algorithm For Word Sense Disambiguation. International Journal on Artificial Intelligence Tools 10(1-2).
  - Omelayenko 2001. **Learning of ontologies** for the Web: the analysis of existent approaches. In Proceedings of the International Workshop on Web Dynamics.
  - Schütze ja Pedersen 1995. **Information Retrieval** Based on Word Senses. In Proceedings of 4th Annual Symposium on Document Analysis and Information Retrieval.

SUMO

käsiteobjektien
hierarkia (isA-suhde)

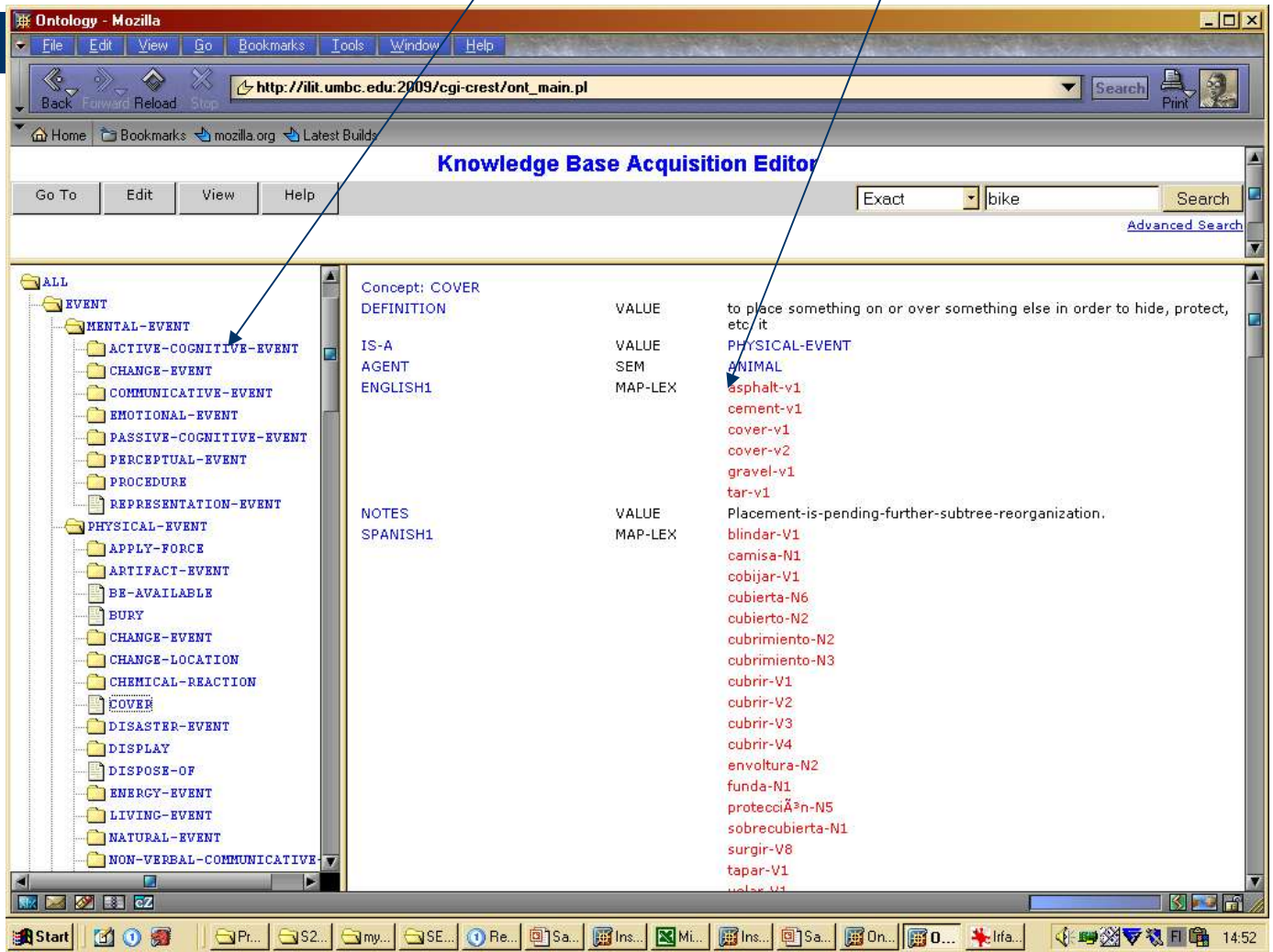
ominaisuudet/piirteet/
assosiaatiot/relaatiot



käsitehierarkia (EVENT,OBJECT,PROPERTY)

käsitteiden ominaisuudet (mukaanlukien espanja-englanti-leksikko)

MikroKosmos



Yhteenveto resursseista