



UiO : Universitetsbiblioteket

Tesaurus-mapping prosjektmøte 10. nov2014
Are Gulbrandsen

Semantikk, pragmatikk og kontekst

- Bruk av vektorromsmodellen for å ta hensyn til kontekst ved mapping**



Vi mapper begreper, ikke termer

- Noen vanlige utfordringer:
 - Synonymer (ulik term for samme begrep)
Kan finnes samme begrep i begge vokabularer, men ulik foretrukken term, f.eks.
'Biologisk mangfold' vs 'Biodiversitet'
 - Homonymer (samme term for to ulike begrep)
 - Ulike bøyningsformer

Semantikk i språkvitenskap (lingv.)

Semantikk $\approx$ Betydning

- Betydningsmessige relasjoner mellom ord
 - To ord med lik betydning = **synonymi**
 - To ord med samme uttrykksform, men helt forskjellig betydning = **homonymi**
 - Forskjellig, men beslektet betydning kalles **polysemi** (gjerne overført betydning (metafor) som tre, stjerne)
- Samme setning kan brukes eller tolkes forskjellig avhengig av situasjonen eller konteksten
- *Studiet av setningers betydning i kontekst* kalles gjerne **pragmatikk**

Semantikk i informasjonsvitenskap

- Hvis syntaksen (kodingen) og semantikken (meningen) for dataene er kjent kan dataene tolkes som informasjon.
- XML og databaser er f.eks. ikke semantisk teknologi
 - Elementnavn/feltnavn er korte tekstlige beskrivelser (termer), som ikke er definerte (i den forstand at man kan bestemme om samme elementnavn i to dokumenter med ulikt skjema faktisk har samme betydning)
 - Er avhengig av situasjonsbestemt tolkning
 - Har bl.a. utfordringer rundt synonymer og homonymer

Semantikk i praksis

- Det er store utfordringer i elektronisk samhandling og kobling av informasjon hvis man ser nærmere på betydning og definisjon av termer som brukes for begreper. – Noen ganger er små forskjeller viktige.
 - Eks: Kungsbacka-saken i Sverige
 - Ca. 1000 personer ble politianmeldt for mistanke om trygdemisbruk da myndighetene sammenlignet informasjon i ulike offentlige systemer.
 - Under 2 % av de anmeldte ble domfelt.
 - Et av forholdene som kom frem under etterforskningen var at det fantes mer enn 25 rettslige definisjoner av «inntekt» innenfor ulike deler av offentlig forvaltning.

Semantikken i RDF, OWL og SKOS

- RDF
 - Ting er identiske (har lik betydning) hvis de har samme uri
- OWL
 - Owl:sameAs
 - Vi kan bytte ut identifikatoren og beholde betydningen
- SKOS
 - skos:closeMatch, skos:exactMatch skos:broadMatch, skos:narrowMatch eller skos:relatedMatch
 - For å være generell definerer SKOS med vilje ikke i detalj hva disse relasjonene *betyr*, det må tolkes i kontekst: "Emphasis on minimal ontological commitment"

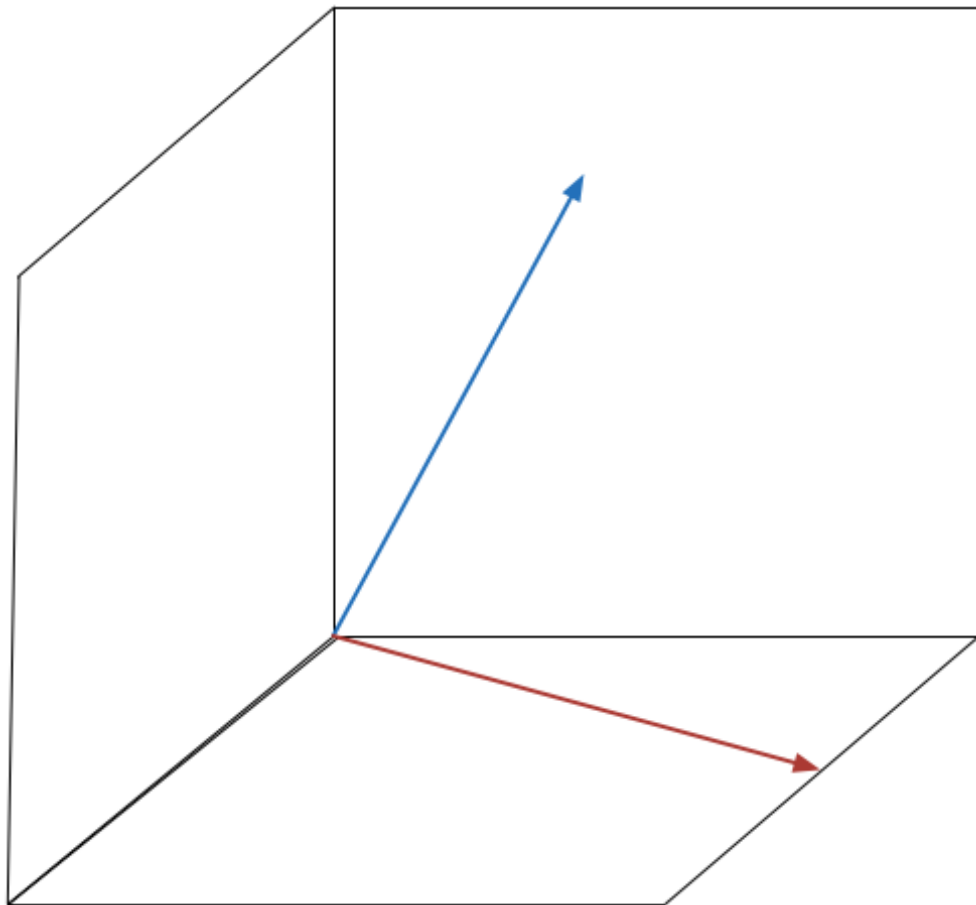
Etablerte metoder og verktøy for mapping

- Mange begrep og metoder i bruk innen litt varierende fagområder, som er overlappende med vår bruk av begrepet mapping, bl.a:
 - «Ontology Mapping», «Ontology Alignment», «Ontology Matching», «Semantic Matching» og «Semantic Mapping»
 - «Linking data without common identifiers»:
 - «Record linkage», «Entity resolution», «Name resolution», «Identity resolution», «Deduplication» og «Merge/purge»

Vektorromsmodellen

- Tar utgangspunkt i statistisk analyse av termer i tekst.
- Et dokument representeres av termenes forekomst og hyppighet i teksten
 - 'Dokument-term'-matrise (tabell)
 - Termvektor
- Likhet representeres og måles med sammenligning av multidimensjonale termvektorer

Likhet basert på vektorvinkel



Kontekst for Emneord i Humord

- I hvilken grad kan ulike former for kontekst gi oss mulighet til å sammenligne humord med Deweyklasser?
 - Overordnede termer
 - Definisjon
 - 'Brukt for'-termer (synonymer)
 - Underordnede termer
 - 'Se også'-termer
 - Noter?

Begreper = term + kontekst

- Begrepet kan representeres av et dokument som består av en vektet sammensetning av bl.a:
 - Term
 - Definisjon
 - Synonymer
 - Overordnede termer
 - Noter

Etablerte verktøy

- Apache Lucene: <http://lucene.apache.org>
 - Støtter Vektorromsmodellen
 - Open Source
 - Modent, og i bruk på UiO
 - Skalerer mer enn bra nok for oss