Luonnollisen kielen tilastollinen käsittely

T-61.281 (3 ov) L

Kevät 2003

Luennot: **Timo Honkela ja Krista Lagus**

Laskuharjoitukset: **Vesa Siivola**

Luentokalvot: Krista Lagus ja Timo Honkela

11. Leksikaalinen semanttinen tieto ja semanttinen samankaltaisuus

Lekseemi: yksittäinen sana-artikkeli sanakirjassa (lexicon).

Leksikaalinen semantiikka: Lingvistinen tutkimusala joka keskittyy lekseemien (sanojen) merkitysten ja niiden sisäisen merkitysrakenteen systemaattiseen tutkimiseen.

Perustavoitteet sanojen merkitysten kuvaamisessa

- 1. Tavoitteena voi olla löytää merkityksen 'perusyksiköt', joiden avulla sanojen merkitykset voitaisiin kuvata.
- 2. Vaihtoehtoisesti voidaan kuvata sanojen väliset merkityssuhteet, esim. merkitysten samankaltaisuus.

Jos ensiksimainittu tunnetaan, toinen seuraa siitä suoraan. Kuitenkin sanojen välisiä merkityssuhteita on mahdollista kuvata vaikka ensiksimainittua eli merkityksen 'perusyksiköitä' ei tunnettaisikaan (vrt. etäisyystaulukko).

Yleisessä kielitieteessä on monia semanttisia teorioita, joita ei kuitenkaan käydä kattavasti läpi tällä kurssilla.

Seuraavassa esitellään Wordnet-tietokanta, jossa on sovellettu yhtä tapaa mallintaa sanoja ja niiden välisiä merkityssuhteita.

Wordnet-tietokanta

- Wordnet on tietokanta, jossa kuvataan englannin sanojen välisiä semanttisia suhteita.
- Pyrkii soveltamaan konsistentisti tiettyä teoriaa sanojen merkitysten representoinnista ja merkityssuhteista.
- Sisältää n. 130,000 sanaa joilla yli 170,000 merkitystä (Wordnet 1.6)
- Tehty käsityönä ihmisvoimin
- Relevanssi tilastollisille menetelmille: tarjoaa tavoitetason merkitysrepresentaatioiden evaluointiin englannin kielelle.

Wordnetin rakenne: substantiivit

Wordnet kuvaa mm. seuraavanlaisia suhteita lekseemien välillä:

Substantiiveille:

Relaatio	Esimerkki
Osana oleminen, osista koostuminen	pöydänjalka → pöytä
Ryhmän suhde jäseneen	tiedekunta → professori
Yläluokat ja aliluokat	ateria → aamiainen
Vastakohdat	vetäjä → seuraaja

Wordnetin rakenne: verbit, adjektiivit ja adverbit

Verbeille:

Relaatio	Esimerkki
Tekemisen kokonaisuus ja osat	nukkua $\rightarrow$ kuorsata
Ylä- ja aliluokat	matkustaa $\rightarrow$ lentää
Vastakohdat	suurentua $\rightarrow$ pienentyä

Adjektiiveille ja adverbeille:

Relaatio	Esimerkki
Vastakohdat	kevyt $\leftrightarrow$ painava
	hitaasti $\leftrightarrow$ nopeasti

Esimerkki yläluokkarelaatiosta

bass, merkitys 3 → laulaja → muusikko → esittävä taiteilija → viihdyttäjä (entertainer) → henkilö, ihminen, joku → elämänmuoto, elollinen organismi, olento → olio, jokin

bass, merkitys 7 → soitin → instrumentti → laite → väline → keinotekoinen esine → fyysinen olio → olio, jokin.

Synonymiteetti

Wordnetissa synonyymeina pidetään sanoja, jotka ovat vaihdettavissa toisikseen tietyssä kontekstissa ilman, että kokonaisilmaisun merkitys oleellisesti muuttuu.

Synset = keskenään tietyssä merkityksessä synonyymisten lekseemien joukko (+ ko. merkityksen kuvaus).

Esim. synset: {*chump, fish, fool, gull, mark, patsy, fall guy, sucker, schlemiel, shlemiel, soft touch, mug*}

Yo. synsetin 'sanakirjamääritelmä': *A person who is gullible and easy to take advantage of*

Polyseemisen (monimerkityksisen) lekseemin merkitykset esitetään kertomalla mihin synsetteihin sana kuuluu, ts. synsetit ovat merkityksen perusyksikköjä. Voidaan ajatella että kukin niistä vastaa jotain *käsitettä* (ajattelun yksikkö, concept).

11.1 Sanojen sisäinen rakenne

Temaattiset roolit

Temaattiset roolien määrittely on lähestymistapa verbien argumenttien semanttiseen kuvaamiseen

Esim. mitä yhteistä on verbeillä 'rikkoa' ja 'avata'?

- molempien toiminnan *kohteena* on tyypillisesti jokin **eloton olio**
- molempien toiminnan *tekijänä* on tyypillisesti jokin **elollinen olento** joka **tahtoo aiheuttaa** tietyn tapahtuman (rikkoutuminen, avautuminen), eli kummankin verbin subjektit ovat *agentteja*.

Esimerkkejä temaattisista rooleista

Joitakin temaattisia rooleja ja esimerkkejä (Jurafsky & Martin, kuva 16.9):

Temaattinen rooli	Esimerkki
AGENT	*Tarjoilija* läikytti keittoa.
EXPERIENCER	*Jukalla* on päänsärky.
FORCE	*Hirmumyrsky* repi talojen kattoja irti.
THEME	Vasta Benjamin Franklinin rikottua *jään* ...
RESULT	Ranskan hallitus on rakentanut kokomääräysten mukaisen *timanttisen pesäpallon*
CONTENT	Mona kysyi *"Tapaat ilmeisesti Niinan keskustassa?"*
INSTRUMENT	Hän ryhtyi virvelöimään ja nosti saaliit veneeseen *haavilla*
BENEFICIARY	Aina kun sihteeri teki pöytävarauksen *pomolleen*, ...
SOURCE	Lensin tänne eilen *Joensuusta*.
GOAL	Ajoin *Jyväskylään*.

Huom. lista ei ole tietyn teorian mukainen eikä ainoa mahdollinen.

Temaattisten roolien hyödyntäminen

Temaattisia rooleja voidaan käyttää matalan tason semanttisessa tulkinnassa esimerkiksi taggaamaan niillä sanoja. Jotkut jäsentimet tuottavat myös tämänkaltaisia semanttisia analyyseja.

Verbien argumentinvalintapreferenssit (selectional preferences)

Useimmat verbit suosivat tiettyyn semanttiseen kategoriaan kuuluvia argumentteja. Tätä ilmiötä kutsutaan nimellä 'selectional preferences' tai 'selectional restrictions' (jos tarkoitetaan kovia sääntöjä).

Esim. *ajatella*-verbin tekijä on tavallisesti ihminen (toisinaan myös eläin tai jokin muu elollinen olento), ei kuitenkaan eloton olio (esim. kirja).

Kuitenkin metaforisessa käytössä voidaan käyttää sanoja toisin kuin yleensä. Esim. 'syödä'-verbi preferoi tekemisen kohteeksi ruoka-argumentteja, mutta: 'hän söi sanansa', 'koira söi kotitehtäväni', 'tuo tapahtuma söi pääministerin valtaa hallituksessa'.

Voidaan myös ajatella että syödä-verbillä on kaksi erillistä käyttötilannetta, konkreettinen ruoan tai objeltin syöminen ja abstrakti. Abstrakti merkitys lainaa konkreettiselta joitakin prosessin ominaisuuksia, kuten syötävän asian väheneminen ja/tai radikaali muodonmuutos syömisen seurauksena.

Preferenssit antavat tietoa tuntemattomista sanoista

'Susan ei ollut koskaan aiemmin syönyt tuoretta duriania'.

Syödä-sanan konkreettisen merkityksen ja siihen liittyvien preferenssien perusteella voisi siis pitää todennäköisenä että 'durian' on ruoka-aine. Vastaavasti 'tuore'-sanan preferenssit viittaavat samaan suuntaan, ja lisäksi siihen että kyse on potentiaalisesti pilaantuvasta ruoka-aineesta.

Eräs menetelmä argumentinvalintapreferenssien estimointiin (Resnik, 1993, 1996)

Tehtävä: selvittää kuinka vahvasti verbi preferoi tiettyjä sanoja (tai sanaluokkia) tietyssä argumentissaan.

Yksinkertaistus 1: Keskitytään tarkastelemaan pelkästään tekemisen kohdeargumentin pääsanaa. Esim. 'Susan söi vihreän **omenan**.' (Jätetään siis huomiotta objektia luonnehtivat tai tarkentavat sanat, tässä 'vihreän').

Yksinkertaistus 2: Yksittäisten substantiivien sijaan tarkastellaan substantiivikategorioita.

Valintapreferenssin voimakkuus

Määritellään **valintapreferenssin voimakkuus** $S(v)$ verbille v.

$$S(v) = D(P(C|v)||P(C)) = \sum_C P(c|v) \log \frac{P(c|v)}{P(c)} \quad (1)$$

$S(v)$ on etäisyys luokkien C priori- ja posteriorijakaumien välillä, laskettuna Kullback-Leibler -divergenssiä (suhteellinen entropia) käyttäen.

Substantiivikategoriat voidaan tuottaa klusteroimalla tai ottaa jostain valmiista leksikaalisesta resurssista (esim. Wordnet).

$S(v)$ kuvaa sitä kuinka valikoiva verbi v on argumenttiensa substantiivikategorioiden osalta.

Valinta-assosiaation voimakkuus

Määritellään seuraavaksi **valinta-assosiaation voimakkuus** $A(v, c)$ verbin v ja kategorian c välillä:

$$A(v, c) = \frac{P(c|v) \log \frac{P(c|v)}{P(c)}}{S(v)}$$

Lopuksi, assosiaation voimakkuus verbin v ja substantiivin n välillä on

$$A(v, n) = \max_{c \in classes(n)} A(v, c)$$

(n voi kuulua useaan substantiivikategoriaan, mikäli on polyseeminen).

Esimerkkejä

Positiiviset $A : n$ arvot kuvaavat positiivisiä preferenssejä, negatiiviset negatiivisia ja 0 on neutraali. Esim.
$A(eat, \text{food}) = 1.08$
$A(find, \text{action}) = -0.13$

Eli syödä-sana preferoi ruokaobjekteja kun taas löytää-sana karkottaa toimintatyyppisiä objekteja. Osa kirjan taulukosta 8.6. (perustuu kategorioihin):

Verbi v	Subst. n	$A(v,n)$	Subst n	$A(v,n)$
answer	request	4.49	tragedy	3.88
find	label	1.10	fever	0.22
hear	story	1.89	issue	1.89
remember	reply	1.31	smoke	0.20
read	article	6.80	fashion	-0.20

Tällaista tietoa voidaan käyttää lauseen semanttisen rakenteen hahmottamisessa, tai esimerkiksi jäsennysten disambiguoinnissa.

11.2 Eräs $S(v)$:n sovellustapa

$S(v)$ ennustaa englannilla melko hyvin sitä, salliiko kyseinen verbi suoran objektin poisjättämisen. Esimerkiksi 'Tiina söi omenaa.' ja 'Tiina söi.' mutta ei niin luontevasti 'Tiina etsi.'

11.3 Semanttinen samankaltaisuus

- Yleispätevän semanttisen representaation löytäminen on kieliteknologian tärkeä ja hankala ongelma.
- Riittävää representaatiota, joka olisi automaattisesti opittavissa ja hyödy ei vielä ole kehitetty.
- Huomattavasti helpompaa on tutkia sanojen semanttista samankaltaisuutta.
- Samankaltaisuutta voidaan hyödyntää esim. ennennäkemättömiin tapauksiin yleistämisessä.

Samankaltaisuuksiin perustuva yleistäminen

Esimerkki: Voiko 'durian' olla syödä-verbin kohde?

Oletetaan, että ei tiedetä mitä 'durian' on, mutta tiedetään, että sen kanssa samankaltaisimmat sanat ovat 'omena', 'banaani' ja 'mango'. Koska kaikki ovat syödä-verbin kohteeksi kelpaavia, voidaan olettaa että myös 'durian' on.

Samankaltaisuuteen perustuva yleistäminen on sukua luokkatietoon perustuvalle yleistämiselle.

Erona on se, että ei ole mitään tiettyä joukkoa viiteryhmiä, joiden sisällä yleistäminen tapahtuu, vaan kulloinenkin viiteryhmä ovat kyseenäolevan yksilön kanssa samankaltaisimmat yksilöt.

Menetelmiä

- **K:n lähimmän naapurin luokitusmenetelmä (KNN):**

  Luokitellaan tuntematon näyte sen perusteella mitkä ovat k:n sitä lähimpänä olevan luokitellun näytteen luokat.

- **Kontekstuaalinen vaihdettavuus semanttisen samankaltaisuuden kriteerinä**

  Miller & Charles (1991): sanojen semanttisen samankaltaisuuden arviot korreloivat vahvasti sanojen vaihdettavuuden kanssa tietyssä kontekstissa.

Monitulkintaisuuden tuoma ongelma: Semanttinen samankaltaisuus on tavallisesti *tiettyjen sananmerkitysten* välinen relaatio, ei polyseemisten sanojen kaikkien merkitysten välinen.

11.4 Vektorietäisyyksiin perustuvat mitat

- **Dokumentti-sana -matriisissa** alkion i, j arvo on 1 jos sana j esiintyy dokumentissa i. 1:n sijaan voidaan käyttää myös sanan lukumäärää.

  Matriisin rivit ovat tällöin *dokumenttivektoreita sana-avaruudessa* ja sarakkeet *sanavektoreita dokumenttiavaruudessa*.

- **Sana-sana -matriisissa** raportoidaan kuinka monta kertaa sana j esiintyi sanan i kontekstissa.

  Konteksi voidaan määritellä eri tavoin, esim. koko dokumentti, kappale, lause, tai jonkin levyinen sanaikkuna.

- Voidaan myös kerätä esim. substantiivien osalta tietoja siitä, mitkä sanat niitä määrittävät (modifier-head-relaatio, määrittäviä sanoja esim. pieni, juokseva,...). Kaksi substantiivia ovat tällöin samankaltaisia siinä määrin kuin niitä määrittävät samat sanat. Kaksi määrittävää sanaa taas ovat samankaltaisia siinä määrin kuin ne määrittävät samoja substantiiveja.

11.5 Esimerkkejä

Dokumentti-sana -matriisin pohjalta laskettaessa esimerkiksi sanat 'astronautti', 'kosmonautti' ja 'kuu' voivat olla samankaltaisia, koska ne kaikki liittyvät avaruusmatkoihin, eli ovat *aihepiiriltään* samankaltaisia.

Sen sijaan katsottaessa lyhyttä kontekstia, tai vielä tarkemmin, tiettyä relaatiota, saadaan tulokseksi toisenlaisia samankaltaisuuksia.

Tällöin 'astronautti' ja 'kosmonautti' olisivat edelleen keskenään samankaltaisia, mutta eivät 'kuun' kanssa—ilmiselvästi koska kuu on taivaankappale ja edellämainitut ihmisiä, joten niitä eivät tyypillisesti määritä samat sanat.

[ks. myös kirjan kuvat 8.3-8.5.]

Reaalilukuvektorien väliset samankaltaisuusmitat

n-ulotteiset reaalilukuvektorit $\mathbf{x}, \mathbf{y} \in R^n$.

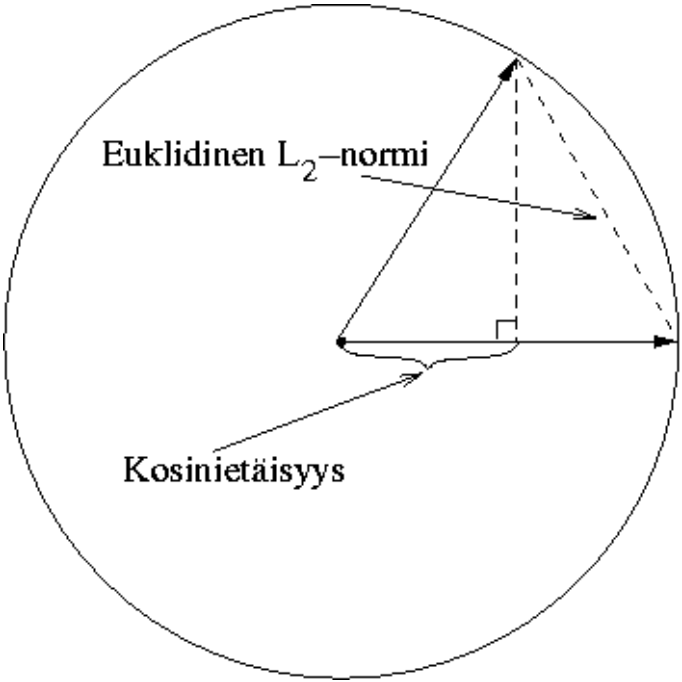
Vektorin pituus on $|\mathbf{x}| = \sqrt{\sum_{i=1}^{n} x_i^2}$

Kosinietäisyys: $cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{|\mathbf{x}||\mathbf{y}|}$, jossa $\mathbf{x} \cdot \mathbf{y}$ tarkoittaa pistetuloa, eli sisätuloa, eli elementeittäin kertomista.

Huom: normalisoiduille (yksikkövektorin pituisille) vektoreille kosinietäisyys on pelkästään vektorien pistetulo. Normalisointi tarkoittaa siis vektorin itsensä pituudella jakamista

Euklidinen L_2-etäisyys: $|\mathbf{x} - \mathbf{y}| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$

Huom: Jos vektorit ovat normalisoituja, ja jos samankaltaisuusmittaa käytetään pelkästään sanojen (tai dokumenttien) järjestämiseen samankaltaisuusjärjestykseen, saadaan Euklidisella ja kosinietäisyydellä aikaan sama järjestys.


Euklidinen L_2–normi
Kosinietäisyys

Esimerkki

Osa kirjan taulukosta 8.8 on esitetty alla. Mukana on tarkasteltava sana ja samankaltaisimmat kosinimitalla laskettuna sana-sana -matriisista.

garlic	sauce	.732	pepper	.728	salt	.726
fallen	fell	.932	decline	.931	rise	.930
engineered	genetically	.758	drugs	.688	research	.687
Alfred	named	.814	Robert	.809	William	.808
simple	something	.964	things	.963	You	.963

Joitain tuloksia

Jos konteksti on suhteellisen pitkä (esim. 25 sanaa kumpaankin suuntaan), tyypillisiä tuloksia:

- joillekin, etenkin keskiharvinaisille sanoille hyviä tuloksia
- aihepiirisamankaltaisuudet keskeisessä roolissa (koska konteksti on melko pitkä)
- sanan eri sananmuodot (esim. verbin preesens ja imperfekti) keskenään samankaltaisia
- ei tunnisteta sanan syntaktista tai semanttista roolia (koska konteksti niin pitkä, ja koska sanojen järjestyksellä ei merkitystä)
- yleisille sanoille, esim. 'simple', saadaan hyvin sekalaisia tuloksia (koska voivat esiintyä hyvin monenlaisissa pidemmissä ts. aihepiirikonteksteissa)

Lisähuomioita

Hyvin lyhyttä kontekstia käytettäessä, esim. vain edeltävä ja seuraava sana, korostuvat puolestaan erityyppiset samankaltaisuudet, esim. sanan syntaktinen ja/tai semanttinen rooli.

Lopuksi: Saadut samankaltaisuudet ovat aina *erittäin* voimakkaasti riippuvaisia valitusta korpuksesta. Ei siis saada yleispäteviä tuloksia.

Vektorietäisyysmittojen hyviä puolia

- Yksinkertaisia ja intuitiivisia, saatu lukuarvo on helposti samaistettavissa samankaltaisuuden asteeseen.
- Mitoilla laskeminen on hyvin nopeaa (etenkin pistetulo, eli kosini normalisoiduilla vektoreilla)
- Mittoja on sovellettu menestyksekkäästi ja pitkään mm. tiedonhaussa.
- Havaittu samankaltaisuutta psykologisen priming-efektin kanssa.

Priming-efekti

Psykologiassa tutkittava ilmiö: sanan havaitsemisen nopeuteen vaikuttaa se, onko lähihistoriassa havaittu ko. sanan kanssa semanttisesti samankaltaista sanaa: jos lähihistoriassa havaittiin esim. sana 'hoitaja' myöhemmin sanan 'lääkäri' havaitseminen tapahtuu nopeammin.

Tutkimalla priming-efektin vahvuutta sanaparien välillä saadaan niille psykologinen samankaltaisuusmitta, jota voidaan periaatteessa käyttää verrokkina yritettäessä kehittää sanojen samankaltaisuus tekstiaineistojen perusteella.

Tällä tavoin saatava tieto on kuitenkin yleensä hyvin työlästä ja hidasta tuottaa, joten tuloksena olevat samankaltaisuusmatriisit ovat hyvin harvoja.

Vektorimittojen huonoja puolia

- Euklidisen metriikan olettaminen ei ole matemaattisesti optimaalista, kun kyse on pohjimmiltaan *lukumäärätiedosta*.
- Esim. etäisyys 0 ja 10 esiintymän välillä on yhtä suuri kuin 990 ja 1000 esiintymän välillä. Jälkimmäinen ero on kuitenkin huomattavasti vähämerkityksisempi.
- Esimerkiksi kosinietäisyys sopii hyvin normaalijakautuneille suureille, mutta ei niin hyvin vinoille jakaumille, kuten lukumäärien jakaumille.
- Mitoilla ei selkeää todennäköisyystulkintaa.

11.6 Todennäköisyyksiin perustuvat mitat

Yhteisesiintymämatriisista saadaan ehdollisten todennäköisyyksien matriisi (tai siis sen ML-estimaatti) jakamalla kunkin alkion lukumäärä ko. rivin alkioiden summalla.

Semanttista samankaltaisuutta (tai pikemminkin erikaltaisuutta) mitataan tämän jälkeen tutkimalla kahden tn-jakauman välistä eroa:

Mitan nimi	Määritelmä
KL divergenssi	$D(p\|\|q) = \sum_i p_i \log \frac{p_i}{q_i}$
Informaatioradius	$D(p\|\|\frac{p+q}{2}) + D(q\|\|\frac{p+q}{2})$
L_1 normi	$\sum_i \|p_i - q_i\|$

(Olet. $0 \log 0 = 0$)

Kullback-Leibler -divergenssi, KL-divergenssi

$D(p||q) = \sum_i p_i \log \frac{p_i}{q_i}$

- Tulkinta: Mittaa paljonko informaatiota kadotetaan, jos oletetaan jakauma q, kun todellinen jakauma on p
- Ongelma 1: vastauksena ∞, jos on yksikin dimensio, jossa $q_i = 0$ ja $p_i \neq 0$ (näin käy usein, jos käytetään ML-estimaatteja todennäköisyyksil
- Ongelma 2: epäsymmetrinen, siis ei ole *metriikka*

Informaatiosäde (IRad)

$D(p||\frac{p+q}{2}) + D(q||\frac{p+q}{2})$

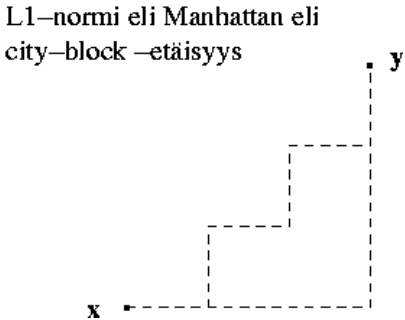
- Paljonko informaatiota kadotetaan, jos kuvataan sekä p että q niiden keskiarvojakaumalla?
- Ei edellisen mitan ongelmia (on symmetrinen, ei ääretöntä arvoa, jos vain jompikumpi tn on nolla)

Kuitenkin: jos todella haluttaisiin ilmaista, että on joitain mahdottomia asioita, niin niitä vastaisi etäisyyksissä ääretön.

L_1 **normi eli Manhattan-etäisyys eli city-block-etäisyys**

$\sum_i |p_i - q_i|$

- Toisistaan poikkeavien tapahtumien suhteellisen frekvenssin odotusarvo.



Loppuhuomioita

- Tn-pohjaisilla mitoilla on selkeä jakaumatulkinta
- Saadut erikaltaisuusmitat täytyy kuitenkin erikseen muuntaa samankaltaisuusmitoiksi (ei aivan suoraviivaista, sisältää parametrin optimoinnin)
- Sekä tn-pohjaiset että vektoripohjaiset mitat hyödyllisiä leksikaalisen tiedon keräämisessä.