Luonnollisten kielten tilastollinen käsittely

T-61.281 (3 ov) L

Kevät 2002

Luennot: **Krista Lagus**
Laskuharjoitukset: **Vesa Siivola**

1. YLEISTÄ KURSSISTA

1.1 Kurssin suorittaminen

Kurssi suoritetaan tekemällä harjoitustyö ja läpäisemällä tentti.

1.2 Ilmoittautuminen

Ilmoittautukaa kurssille www-topin avulla. Kieliteknologian opetuksen verkoston (KIT) opiskelijat voivat ilmoittautua sähköpostitse kurssin luennoijalle, kunnes saavat opintokirjannumeron TKK:lle.

1.3 Tiedotukset

Kurssista tiedotetaan webissä http://www.cis.hut.fi/Opinnot/T-61.281, ryhmässä news://nntp.tky.hut.fi/opinnot.tik.informaatiotekniikka sekä CIS-laboratorion ilmoitustaululla 3.krs aulassa B-käytävän suulla.

1.4 Luennot

Luennot pidetään keskiviikkoisin kello 10–12 salissa T2.

Luennoitsija opettava tutkija TKT Krista Lagus (mailto:krista.lagus@hut.fi).

Luentokalvot ovat luennon jälkeen nähtävillä osoitteessa http://www.cis.hut.fi/Opinnot/T-61.281/kalvot.pdf.

Luennoitsijan vastaanotto on tarvittaessa välittömästi luennon jälkeen T-talossa huoneessa C310.

1.5 Laskuharjoitukset

Laskuharjoitukset tiistaisin kello 8–10 salissa U264 alkaen 22.1.2001.

Harjoitukset pitää DI Vesa Siivola (mailto:vesa.siivola@hut.fi).

Tehtävät laitetaan edellisenä perjantaina nähtäville osoitteeseen http://www.cis.hut.fi/Opinnot/T-61.281/laskarit.html.

1.6 Kirja

Kurssi seuraa kirjaa:

Christopher D. Manning, Hinrich Schütze:
Foundations of statistical natural language processing,
MIT Press, 1999.

Kirja löytyy TKK:n pääkirjastosta ja tietotekniikan kirjastosta.

Yliopiston kirjakauppa (TKK) tilaa kirjaa haluttaessa sitovilla etukäteistilauksil
(hinta n. 600 mk).

Kirjan hinta muutamassa web-kaupassa:
http://www.uk.bol.com/ noin 74 euroa (sis. kuljetus)
http://www.amazon.com/ noin 87 euroa (sis. kuljetus)
http://www.amazon.co.uk/ noin 79 euroa (sis. kuljetus)

Tutustumiskappale on nähtävillä laboratorion sihteerin Tarja Pihamaan huo-
neessa B326 olevassa harmaassa peltisessä vetolaatikostossa.

1.7 Luentomonisteet

Laskuharjoitukset ratkaisuineen ja luentokalvot ilmestyvät opetusmonisteina.

Materiaalia voi myös lainata luennoitsijalta ja assarilta itse kopioitavaksi.

Vapaaehtoinen kurssitoimittaja?

1.8 Suhde muihin opintoihin

Kurssi soveltuu osaksi seuraavia opintoja

- Kieliteknologian pää- ja sivuaine TKK:lla (Tik, Sähkö)
- Informaatiotekniikan pää- ja sivuaineen valinnaiset opinnot
- KIT-verkoston opinnot (mm. Helsingin Yliopistolla)
- Muut aiheeseen liittyvät jatko-opinnot TKK:lla ja muualla (hyväksytettä
  erikseen)

1.9 Tentti

Tentti järjestetään 14.toukokuuta klo 13-16. Lisäksi syksyn tenttikausilla järjestetään yksi tai kaksi tenttiä.

Tentissä saa olla mukana matemaattinen kaavakokoelma ja tavallinen funktiolaskin.

Tenttiin ilmoittaudutaan normaalisti eli Topin kautta viimeistään 2 päivää etukäteen.

1.10 Harjoitustyö

Kurssin suoritukseen kuuluu pakollinen harjoitustyö.

Harjoitustyön tehtävänanto, arvostelu ja aiheet esitellään luennolla myöhemmin keväällä, jolloin aiheet laitetaan myös esille osoitteeseen http://www.cis.hut.fi/Opinnot/T-61.281/harjtyo.html.

2. JOHDANTO

2.1 Mitä on tilastollinen luonnollisen kielen käsittely?

- Kieliteknologian osa-alue
- Sovelletaan informaatiotekniikan, tilastomatematiikan, ja tietojenkäsittelytieteen menetelmiä kieliteknologisiin ongelmiin.
- Malliperheet sisältävät todennäköisyyksiä, jotka estimoidaan hyvin suurista aineistoista (*korpuksista*).
- Sovelluskohteita: tiedonhaku, tekstien järjestäminen ja luokittelu, puheentunnistus, luonnollisen kielen käyttöliittymät tietokantoihin ja varauspalveluihin
- Menetelmäaloja: koneoppiminen, hahmontunnistus, tilastotiede, todennäköisyyslasku, signaalinkäsittely
- Lähialoja: lingvistiikka, korpuslingvistiikka, fonetiikka, keskusteluntutkimus, tekoälyntutkimus, kognitiotiede

- Nykykäsitys: Lingvististä tietoa kielestä hyödynnetään prioritietona, sen sijaan että lähdettäisiin täysin puhtaalta pöydältä.

2.2 Mallinnuksen peruskäsitteitä

- Malli — Jonkin ilmiön tai datajoukon kattava kuvaus.
  Esim: sääntökokoelma joka kuvaa suomen morfologian.
- Malliperhe, malliavaruus — joukko potentiaalisia malleja joita harkitaan ilmiön kuvaamiseen. Esim. niiden sääntöjen kokoelma jota voitaisiin periaatteessa käyttää kielen syntaksin kuvaamiseen.
- Mallin valinta — prosessi jonka kautta päädytään johonkin tiettyyn malliin. Algoritmit usein tämäntyyppisiä: vuorotellaan mallin evaluointia ja mallin muuttamista, pyrkien kohti parempaa mallia.
- Oppiminen — ks. mallin valinta.
- Probabilistinen malli(perhe) — representoi ilmiöiden todennäköisyyksiä.
- Algoritmi — menetelmä jolla edetään malliavaruudessa, kokeillen eri malleja.
- Iteratiivinen — vähän kerrassaan, toiston kautta tapahtuva

2.3 Yleisestä kielitieteestä

Tavoiteena kuvata ja selittää toisaalta kielen (kielten) säännönmukaisuudet, toisaalta kielen (kielten) monimuotoisuus.

Oppivien menetelmien terminologialla ilmaistuna: tavoiteena on *konstruoida malli kielestä*.

Kielen ilmenemismuotoja mm. keskustelut visuaalisella kontaktilla ja ilman, viitomalla, yksinpuhelut, kirjoitetut artikkelit, kirjat, luennot, ja muut kielelliset viestit eri viestinvälineitä ja -ympäristöjä käyttäen.

Laajemmin nähtynä kielen mallinnuksen tavoitteena on selvittää ja kuvata:

- Miten ihmiset käyttävät kieltä, mitä todella sanotaan?
- Mitä kielenkäyttäjä tahtoo tai mihin pyrkii sanoessaan jotain?

Kuvattavan ilmiön osa-alueita:

- Kogitiiviset mekanismit, eli miten kielikyky syntyy ja muotoutuu ihmisessä (ja muissa olennoissa), ja miten tuotamme ja ymmärrämme kieltä.
- Kielen ilmausten ja maailman väliset yhteydet (entä *maailmantiedon* kuvaus?)
- Kielessä esiintyvät säännönmukaisuudet.

2.4 Perinteinen lähestymistapa kielitieteessä

Ominaisuus 1: Perinteisen lähestymistavan mukaan kieli on kuvattavissa *joukkona 'kovia' sääntöjä*, esim. produktiosääntöjä.
<= MALLIPERHE

Esimerkki: Englannin substantiivilauseke NP koostuu valinnaisesta artikkelista DET=[a, the, an], valinnaisesta määrästä adjektiiveja ADJ=[brown, beautiful,...] ja substantiivista N=[flower, building, thought...].

NP => (Det)? (ADJ)* N

Ominaisuus 2: Sääntöjen avulla pyritään kuvaamaan mitkä lauseet ovat hyvinmuodostettuja (sallittuja, kieliopin mukaisia) ja mitkä väärinmuodostettuja (kiellettyjä, kieliopin vastaisia).
<= MALLINNUKSEN TAVOITE

Mallinnuksen tavoite aina kahtalainen: *kattavuus* ja *tarkkuus*.

'Kaikki kieliopit vuotavat' (Edward Sapir, 1921)

Täydellisen kuvauksen saavuttamisen esteinä ainakin kielellinen variaatio (yksilöiden ja kieliyhteisöjen välillä), luovuus, kielen muuttuminen.

Kritiikki 1: Onko kovan kieliopillinen-eikieliopillinen rajan etsiminen hyvin määritelty ongelma, ts., onko sellaista rajaa edes olemassa, vai onko kyse aidosti sumeasta ilmiöstä?

On paljon lauseita joiden kieliopillisuudesta voidaan olla montaa mieltä, ja ollaankin. $=>$ Todellisuudessa kovaa rajaa ei ehkä ole.

Kritiikki 2: Onko kieliopillisuus relevantti ja riittävä kielen kuvauksen taso?

Esim. lause 'Colourless green ideas sleep furiously.' (Chomsky) on syntaktisesti ok, mutta semanttisesti ei kovin mielekäs tai ainakaan tavanomainen.

Ratkaisuyritys: määritellään myös semanttisia sääntöjä. Ongelmia kuitenkin tulee, mm. sanojen metaforisen käytön kanssa. Ehkä 'kovat' säännöt ylipäänsä eivät ole oikea malliperhe?

Esimerkki:

Sääntö: niellä-sanan subjektina täytyy olla elävä olento
Lause: Supernova nielaisi planeetan.

2.5 Kategoriset (diskreetit) vs. jatkuvat representaatiot

- a/ä p/b: äänisignaalissa jatkuva muutos, foneemitasolla havainto on kategorinen: havaitaan joko a tai ä, ja havainto muuttuu yhtäkkisesti jossain kohti signaalin muuttuessa vähitellen.

Havainttaessa puhetta muutos jatkuvalta representaation tasolta (äänisignaali) diskreetiksi tai kategoriseksi (foneemi). Puhetta tuotettaessa päinvastainen muutos.

Todellisissa systeemeissä eroa diskreetin ja jatkuvan välillä ei ole, koska:

- Kaikki todelliset systeemit ovat kohinaisia (fysiikan perusteet)
- kohinainen kommunikaatikanava aina diskretoi signaalin (Shannonin informaatiteoria)

Sen sijaan aidosti relevantti kysymys on, onko representaatioavaruuden pisteiden välille määritelty etäisyysrelaatio (metriikka) vai ei. Usein tarkoitetaankin tätä silloin kun puhutaan jatkuvista representaatioista.

2.6 Probabilistinen esitystapa

- Probabilistisessa mallissa malliperheenä todennäköisyydet. (vertailukohta: kaksiarvoinen esitystapa jossa asiat ovat joko-tai, tosia tai epätosia
- Esitystapa mahdollistaa tiedon esittämisen silloinkin kun ei voida muodostaa kategorista sääntöä, mutta on olemassa preferenssi: Subjekti on ennen predikaattia 90% tapauksista P(A)=0.9.
- 'Kova' sääntö: P(A)=1 tai P(A)=0.
- Probabilistisessa representaatiossa tiedon kerääminen ja mallin päivittäm voi tapahtua iteratiivisesti, vähitellen. Lisäesimerkit tarkentavat aiemmin muodostettua alustavaa kuvaa.

2.7 Datasta oppiminen

Periaatteellinen lähtökohta mallinnukseen

Mallit eivät ole tosia tai epätosia. Ne ovat parempia tai huonompia. Paremmuus mitataan suurista joukoista todellista dataa kokonaisuutena. Vertailukohta tai onnistumisen mitta ei ole vastaavuus lingvistisen intuition kanssa, vaan ilmiön/datan optimaalinen kuvaus.

Perustelu oppimiselle

Miksi kannattaa muodostaa malleja automaattisesti, datasta oppimalla tai estimoimalla (eli automaattisesti), eikä asiantuntijatietoa kirjaamalla?

- Data on halpaa ja sitä on paljon, myös sähköisesti.
- Voidaan saada mallit aikaan nopeammin / vähemmällä ihmistyövoimalla / pienemmin kustannuksin.
- Kielen muuttuessa mallit voidaan estimoida uudestaan helposti.

- Asiantuntijatietämys hankalaa tuottaa tai kerätä (mm. konsistenssiongelmat).
- Asiantuntijatietoa käytettäessä malliperhettä rajoittaa 'ihmisbias'.
- Koneiden 'kognitiiviset ominaisuudet' eroavat ihmisen vastaavista.
- Toteutettaessa kielikykyä koneille ei tarvitse rajoittua ihmiselle helposti ymmärrettäviin malleihin.
- Aineistolähtöinen keskittää resurssit niihin ilmiöihin jotka todella esiintyvät. Resurssien käyttö suhteessa ilmiön keskeisyyteen aineistossa.

Onnistuneen oppivan mallinnuksen seurauksia

- Resurssien käytön tehostuminen: Voidaan ulottaa mallinnus laajempaan kielijoukkoon, ja yksittäisen kielen sisällä eri osa-alueisiin.
- Laadullinen parannus, koska koneellisesti pystytään käymään läpi suuri joukko malleja ja koska mallin valinnassa ei ole inhimillistä biasta (ainakaan samassa määrin kuin käsin muodostetuissa malleissa).

Riskejä ja haasteita

- Datan valinta ja kattavuus,
- sopivien malliperheiden määrittely,
- optimointimenetelmien tehokkuus.

2.8 Ihmisen kielikyky ja kielen oppiminen

Miten kielikyky ihmisellä syntyy ja muotoutuu? Mikä osa on synnynnäistä, mitä opitaan?

Rationalistinen näkemys: Kielikyky on synnynnäinen, ja oma erillinen kielimodulinsa

Keskeisiltä osin ihmismielen ja kielen rakenne on kiinnitetty (oletettavasti geneettisesti määrätty). Perustelu: argumentti stimuluksen vähyydestä (mm. Chomsky 1986). Kannattajia mm: Chomsky, Pinker.

Vrt. tekoälyntutkimus 1970-luvulla: tietämyksen koodaaminen käsin. Saatiin aikaan pienimuotoisia älykkään oloisesti käyttäytyviä systeemejä (mm. Newell & Simon: Blocks world). Systeemit usein käsin koodattuja sääntöpohjaisia järjestelmiä. Näiden laajentaminen on kuitenkin osoittautunut hyvin hankalaksi.

Empiristinen näkemys: Kieli opitaan, kielikyky toteutuu osana yleistä kognitiivista laitteistoa

Amerikkalaiset strukturalistit. Zellig Harris (1951) jne: tavoitteena kielen rakenteen löytäminen automaattisesti analysoimalla suuria kieliaineistoja. Ajatus siitä että hyvä rakennekuvaus (grammatical structure) on sellainen joka kuvaa kielen kompaktisti.

Nykyisin melko yleisen näkemyksen mukaan mieli ei ole täysin tyhjä taulu, vaan oletetaan että tietyt 1. rakenteelliset preferenssit yhdessä 2. yleisten kognitiivisten oppimisperiaatteiden ja 3. sopivanlaisen stimulin kanssa johtavat kielen oppimiseen.

Vrt. adaptiivisten menetelmien tutkimus, havaintopsykologia ja laskennallinen neurotiede, ihmisen havaintomekanismien ja piirreirroittimien muotoutuminen aistisyötteen avulla (*plastisiteetti*).

Avoimia kysymyksiä:

- Tarvittavan prioritiedon määrä ja muoto?
- Mitä ovat tarvittavat oppimisperiaatteet?
- Minkälaista syötettä ja missä järjestyksessä tarvitaan?

Käytännöllinen lähestymistapa

Tavoite voi olla puhtaasti käytännöllinen: kehittää toimivia, tehokkaita kieliteknologisia menetelmiä ja järjestelmiä.

Eri menetelmiä sovellettaessa ei välttämättä oteta rationalismi-empirismi-vastakkainasetteluun lainkaan kantaa.

Aineistoihin (korpuksiin) pohjautuvat ja tietämysintensiiviset mallit ovat tällöin samalla viivalla.

Vertailukriteerit:

- lopputuloksen laatu
- lopullisen mallin tilankäytön tehokkuus ja riittävä nopeus (esim. reaaliaikaiset sovellukset)
- mallin konstruoinnin tai oppimisen tehokkuus (tarvittava ihmistyö, prosessointitila ja -aika)

Usein kohteena jokin spesifi kieliteknologinen sovellusongelma, jonka ratkaisemiseksi riittää vain osittainen kielen mallinnus.

Koko kielikyvyn implementointi luultavasti edellyttäisi koko kognitiivisen väline ja tekoälyn toteuttamista, mukaanlukien maailmantiedon kerääminen ja esittär

3. MATEMAATTISIA PERUSTEITA

3.1 Todennäköisyyslasku

Peruskäsitteitä

Todennäköisyysavaruus (*probability space*):
Tapahtuma-avaruus Ω — diskreetti tai jatkuva
Todennäköisyysfunktio tai -jakauma P
Kaikilla tapahtuma-avaruuden pisteillä A on todennäköisyys: $0 \leq P(A) \leq 1$

Todennäköisyysmassa koko avaruudessa on $\sum_A P(A) = 1$

Esimerkki 1

Jos tasapainoista kolikkoa heitetään 3 kertaa, mikä on todennäköisyys että

saadaan 2 kruunaa?

Mahdolliset heittosarjat Ω: { HHH, HHT, HTH, HTT, THH, THT, TTH, TTT }

Heittosarjat joissa 2 kruunaa: A={ HHT, HTH, THH}

Oletetaan tasajakauma: jokainen heittosarja yhtä todennäköinen, $P = 1/8$

$$P(A) = \frac{|A|}{|\Omega|} = \frac{3}{8}$$

3.2 Ehdollinen todennäköisyys

A= asiaintila jonka todennäköisyyden haluamme selvittää
B= meillä oleva ennakkotieto tilanteesta, ts. tähän asti tapahtunutta

Ehdollinen todennäköisyys, A:n todennäköisyys ehdolla B:

$$P(A|B) = \frac{P(A,B)}{P(B)} \tag{1}$$

Palataan esimerkkiin 1: Oletetaan että on jo heitetty kolikkoa kerran ja saatu kruuna. Mikä nyt on todennäköisyys että saadaan 2 kruunaa kolmen heiton sarjassa?

Alunperin mahdolliset heittosarjat: {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}

Prioritiedon B perusteella enää seuraavat sarjat mahdollisia: { HHH, HHT, HTH, HTT }

$P(A|B) = 1/2$

Ketjusääntö

$$P(A_1, ...A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2)\dots P(A_n|A_1, \dots, A_{n-1})$$

Riippumattomuus

Tilastollinen riippumattomuus:

$$P(A, B) = P(A)P(B) \tag{2}$$

Sama ilmaistuna toisin: se että saamme lisätiedon B ei vaikuta käsitykseemme A:n todennäköisyydestä, eli:

$P(A) = P(A|B)$

Huom: tilastollinen riippuvuus $\neq$ kausaalinen riippuvuus!

Esim. jäätelön syönnin ja hukkumiskuolemien välillä on tilastollinen riippu-

vuus. (Yhteinen kausaalinen tekijä ehkä lämmin kesäsää.)

Ehdollinen riippumattomuus

$$P(A, B|C) = P(A|C)P(B|C) \tag{3}$$

A ja B ovat riippumattomia ehdolla C mikäli on niin että jos jo tiedämme C:n, tieto A:sta ei anna mitään lisätietoa B:stä (ja päinvastoin).

3.3 Bayesin teoreema

Koska kyse ei ole kausaalisesta riippumisesta, tapahtumien järjestystä voidaan vaihtaa:

$$P(A, B) = P(B)P(A|B) = P(A)P(B|A) \tag{4}$$

Eli $P(B|A)$ voidaan laskea $P(A|B)$:n avulla.

$$P(B|A) = \frac{P(A, B)}{P(A)} = \frac{P(B)P(A|B)}{P(A)} \tag{5}$$

Jos $A =$ lähtötilanne, joka ei muutu (esim. jo tapahtuneet asiat), ja haluamme ainoastaan tietää, mikä tulevista tapahtumista B on todennäköisin,

$P(A)$ on normalisointitekijä joka voidaan jättää huomiotta:

$$\arg\max_B P(B|A) = \arg\max_B \frac{P(B)P(A|B)}{P(A)} = \arg\max_B P(B)P(A|B) \quad (6)$$

Toisaalta, $P(A)$ voidaan myös laskea:

$P(A) = \sum_i P(A|B_i)P(B_i)$

3.4 Satunnaismuuttuja

Satunnaismuuttuja (*random variable*)

Jatkuva-arvoinen satunnaismuuttuja: $X : \Omega => \mathbb{R}^n$, jossa $\mathbb{R}$ on reaalilukujen joukko ja n on avaruuden dimensio. Jos $n > 1$ puhutaan myös satunnaisvektorista.

Diskreetti satunnaismuutuja: $X : \Omega => S$, jossa S on numeroituva $\mathbb{R}$:n osajoukko.

Indikaattorimuuttuja: $X : \Omega => 0, 1$.

Todennäköisyysjakauma *probability mass function pmf* $p(x)$

3.5 Odotusarvo ja varianssi

Odotusarvolle $E(X) = \sum_x xp(x)$

(äärellisessä tapauksessa; äärettömässä tapauksessa summan korvaa integraali)

Ts. odotusarvo on (painotettu) keskiarvo.

Varianssi kuvaa muuttujan arvon vaihtelua keskiarvon ympärillä:

$$\begin{aligned} Var(X) &= E((X - E(X))^2) \\ &= E(X^2) - E^2(X) \end{aligned}$$