Luonnollisen kielen tilastollinen käsittely

T-61.281 (3 ov) L

Kevät 2004

Luennot: **Timo Honkela**

Luentokalvot: Krista Lagus ja Timo Honkela

1. YLEISTÄ KURSSISTA 1
1.1 Kurssin suorittaminen 1
1.2 Ilmoittautuminen 1
1.3 Tiedotukset 2
1.4 Luennot 3
1.5 Laskuharjoitukset 4
1.6 Kirja 5
1.7 Luentomonisteet 6
1.8 Suhde muihin opintoihin 7
1.9 Tentin järjestelyt 8
1.10 Tenttikysymyksistä 9
1.11 Harjoitustyö 10
2. JOHDANTO 11
2.1 Tilastollinen luonnollisen kielen käsittely 11
2.2 Luonnollisen kielen käsittelyn sovelluskohteita . . . 12
2.3 Mallinnuksen peruskäsitteitä 13
2.4 Yleisestä kielitieteestä 14
2.5 Lähestymistapoja kieli-ilmiöihin 15

2.6 Perinteinen lähestymistapa kielitieteessä 16
2.7 Kielen mallintamisen haasteita 17
2.8 Perinteisen lähestymistavan ongelmia, 1 18
2.9 Perinteisen lähestymistavan ongelmia, 2 19
2.10 Kategoriset (diskreetit) vs. jatkuvat representaatiot 20
2.11 Probabilistinen esitystapa 21
2.12 Probabilistisen esitystavan ja sumean esitystavan suhde 22
2.13 Perusteluja datasta oppimiselle, 1 23
2.14 Perusteluja datasta oppimiselle, 2 24
2.15 Ihmisen kielikyky ja kielen oppiminen 26
3. MATEMAATTISIA PERUSTEITA 31
3.1 Todennäköisyyslasku 31
3.2 Ehdollinen todennäköisyys 33

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.

Luennoitsijat: professori (ma), fil.tri Timo Honkela
ja opettava tutkija, tekn.tri Krista Lagus
(mailto:timo.honkela@hut.fi).

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

Kalvot perustuvat Krista Laguksen vuonna 2002 pitämään kurssiin.

1.5 Laskuharjoitukset

Laskuharjoitusten aika ja paikka käsitellään ke 21.1. luennon yhteydessä.

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.

Tutustumiskappale on nähtävillä laboratorion sihteerin Tarja Pihamaan huonee
B326 olevassa harmaassa peltisessä vetolaatikostossa.

1.7 Luentomonisteet

Laskuharjoitukset ratkaisuineen ja luentokalvot ilmestyvät opetusmonisteina kurssin lopuksi. Laskuharjoituksissa paikan päällä jaetaan mallivastaukset, jotka ovat myös opetusmonisteessa.

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 yliopistossa)
- Muut aiheeseen liittyvät jatko-opinnot TKK:lla ja muualla (hyväksytettävä erikseen)

1.9 Tentin järjestelyt

Tentti järjestetään toukokuussa. Tarkempi ajankohta selviää osaston tenttijärjestyksen valmistuttua. Lisäksi syksyn tenttikausilla järjestetään yksi tai kaksi tenttiä.

Tentissä on 5 tehtävää $\grave{a}$ 6 pistettä, maksimi 30 pistettä.

Tentissä saa olla mukana matemaattinen kaavakokoelma ja tavallinen funktiolaskin.

Tenttiin ilmoittaudutaan normaalisti eli Topin kautta viimeistään 2 päivää etukäteen.

1.10 Tenttikysymyksistä

Tentissä pyritään mittaamaan sitä kuinka hyvin opiskelija on perehtynyt toisaalta tilastollisen kielenkäsittelyn sovellusongelmiin ja toisaalta alan keskeis in menetelmiin.

Tehtävät tulevat painottumaan luentomonisteiden ja laskarien käsittelemiin aiheisiin. Kuitenkin kirjan lukeminen näiden aiheiden osalta on suositeltavaa.

Tehtävät voivat olla esseetehtäviä, pieniä sanallisia tehtäviä ja laskutehtäviä. Laskutehtävät ovat samantyyppisiä kuin laskareissa.

Tehtävinä voi olla esim. tietyn sovellusongelman selostaminen (mistä on kysymys), mitä menetelmiä ongelmaan on käytetty tai voidaan käyttää, jonkin (tietyn) menetelmän selostaminen yksityiskohtaisesti, tai eri menetelmie hyvien ja huonojen puolien vertaaminen.

Voidaan myös edellyttää kykyä tulkita mitä oletuksia jossain mallissa tehdään, ja arvioida kuinka paikkansapitäviä ne ovat ko. sovellusongelman kannalta.

1.11 Harjoitustyö

Kurssin suoritukseen kuuluu pakollinen harjoitustyö.

Jos haluaa kurssista suoritusmerkinnän toukokuun tenttitulosten yhteydessä, harjoitustehtävä on saatava hyväksytysti läpi toukokuun 1. päivään mennessä.

Harjoitustyön tehtävänanto, arvostelu ja aiheet esitellään luennolla kahden viikon kuluttua 28.1.2003,
jolloin aiheet laitetaan myös esille osoitteeseen
http://www.cis.hut.fi/Opinnot/T-61.281/harjtyo.html.

2. JOHDANTO

2.1 Tilastollinen luonnollisen kielen käsittely

- Kieliteknologian osa-alue
- Sovelletaan informaatiotekniikan, tilastomatematiikan, ja tietojenkäsittelytieteen menetelmiä kieliteknologisiin ongelmiin.
- Rakennetaan malleja luonnollisesta kielestä niin, että niiden sisältämät todennäköisyysarvot estimoidaan (hyvin) suurista aineistoista (nk. *korpuksista*).
- Menetelmäaloja: koneoppiminen, hahmontunnistus, tilastotiede, todennäköisyyslasku, signaalinkäsittely
- Lähialoja: kielitiede, korpuslingvistiikka, fonetiikka, keskusteluntutkimus, tekoälytutkimus, kognitiotiede

2.2 Luonnollisen kielen käsittelyn sovelluskohteita

Sovelluskohteita ovat mm.

- tiedonhaku
- tekstien järjestäminen ja luokittelu
- puheentunnistus
- luonnollisen kielen käyttöliittymät
  esimerkiksi tietokantoihin ja varauspalveluihin

2.3 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) — esittää ilmiöiden todennäköisyyksiä.
- Iteratiivinen — vähän kerrassaan, toiston kautta tapahtuva

2.4 Yleisestä kielitieteestä

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

Tavoitteena on *konstruoida malli kielestä*.

Kielen ilmenemismuotoja mm. keskustelut visuaalisella kontaktilla ja ilman, viittomalla, 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 kielen käyttäjä tahtoo tai mihin pyrkii sanoessaan jotain?

2.5 Lähestymistapoja kieli-ilmiöihin

- Autonominen kielitiede:
  Selvitetään kielissä esiintyviä säännönmukaisuuksia ja variaatiota.

- Kognitiivinen kielitiede:
  Selvitetään kielen käsittelyyn liittyviä kognitiivisia mekanismeja, kuten sitä, miten kielikyky syntyy ja muotoutuu ihmisessä (ja muissa olennoissa), ja miten tuotamme ja ymmärrämme kieltä.

- Luonnollisen kielen käsittely tekoälyn osa-alueena:
  Kehitetään kielen ilmausten automaattisen tulkinnan ja tuottamisen mekanismeja. Selvitetään kielen ja maailman välisiä yhteyksiä ja kehitetään malleja niiden toiminnalliseen kuvaukseen.

2.6 Perinteinen lähestymistapa kielitieteessä

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

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 väärinmuodostettu (kiellettyjä, kieliopin vastaisia).

Mallinnuksella on kaksi tavoitetta: *kattavuus* ja *tarkkuus*.

2.7 Kielen mallintamisen haasteita

- Monitulkintaisuudet
- Tulkinnan kontekstuaalisuus
- Kielen sumeus
- Kielen muuttuminen
- Tulkinnassa tarvittavan tietämyksen määrä ja laatu
- Multimodaalinen kommunikaatio
- Tulkinnan subjektiivisuus ja intersubjektiivisuus

2.8 Perinteisen lähestymistavan ongelmia, 1

'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 - ei-kieliopillinen -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.

2.9 Perinteisen lähestymistavan ongelmia, 2

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.

2.10 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.

Havaittaessa 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

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

2.11 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.

2.12 Probabilistisen esitystavan ja sumean esitystavan suhde

- Probabilistinen näkökulma: kuinka todennäköinen jokin tapahtuma on.
- Sumeus: missä määrin jokin alkio kuuluu johonkin joukkoon, tms.

2.13 Perusteluja datasta oppimiselle, 1

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'.

2.14 Perusteluja datasta oppimiselle, 2

- 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.

2.15 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. tekoälytutkimus 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 kielite-knologisia 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. reaali-aikaiset sovellukset)
- mallin konstruoinnin tai oppimisen tehokkuus (tarvittava ihmistyö, pros-essointitila ja -aika)

Usein kohteena jokin spesifi kieliteknologinen sovellusongelma, jonka ratkaisem
si 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 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).