Tiedonhakumenetelmät

Helena Ahonen-Myka

Kevät 2004, osa 8

Ryvästäminen: esimerkki

Luonnollisen kielen käsittely hakua varten

Tässä osassa

- luennolla käydään läpi esimerkki täydellisen linkityksen kriteerin käytöstä ryvästämisessä
- luonnollisen kielen käsittely hakua varten
  - luonnollisen kielen piirteitä
  - luonnollisen kielen käsittelymenetelmiä
  - käänteistiedostoratkaisuja

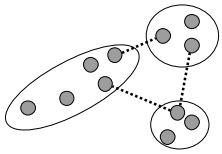
2

Ryvästäminen: täydellisen linkityksen kriteeri

- käydään läpi esimerkki Saltonin kirjan sivuilta 333-336

3

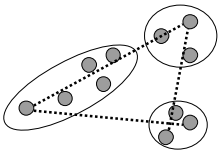
Yhden linkin kriteeri



Yhdistetään ne kaksi ryvästä, joiden lyhin etäisyys on pienin.

4

Täydellisen linkityksen kriteeri



Yhdistetään ne kaksi ryvästä, joiden pisin etäisyys on pienin.

5

Luonnollisen kielen yleisiä ongelmia

- luonnollisen kielen monimuotoisuus käsitteiden ilmaisussa on keskeinen tiedonhaun ongelma
- piirteitä:
  - kieli ei ole yhteistä
  - kieli kehittyy
  - monitulkintaisuus

6

Luonnollisen kielen ominaisuuksia

- **homonymia**: kahden eri sanan kirjoitus- ja äänneasut ovat identtiset
  - homografia: vain kirjoitusasut ovat identtiset
    - "hauissa" (sanan haku tai sana hauki taivutusmuoto)
  - täydellinen homonymia: kaikki sanojen taivutusmuodot ovat identtisiä
  - sanamuotohomografia: osa taivutusmuodoista on identtisiä
  - polysemia: samalla sanalla on kaksi tai useampia lähekkäisiä merkitysvariantteja (esim. "kaula")

7

Luonnollisen kielen ominaisuuksia

- **synonymia:** sama käsite voidaan ilmaista monella eri sanalla
  - synonymiasuhde ei ole aina täsmällinen
  - synonyymeillä saattaa olla yhteinen päämerkitys ja eri sivumerkitykset
- **kiertoilmaukset**
  - käsitteelle ei ole olemassa (tai kirjoittaja/puhuja ei tunne) lyhyttä nimitystä
  - käsite ilmaistaan monimutkaisemman ilmauksen (parafraasin) avulla
  - käsite voi olla myös ilmaistu "rivien välissä"

8

Luonnollisen kielen ominaisuuksia

- **anaforiset ilmaukset:** epätäydellisiä ilmauksia, jotka viittaavat tekstiyhteydessä aiemmin esitettyyn kohteeseen
  - "Valkeakosken paperitehdas lomauttaa työntekijöitä. **Tehtaan** käyttökate on romahtanut."
  - pronominit viittaavat johonkin edeltävään sanaan
- **ellipsi:** vaillinainen ilmaisu, jossa kielellisen ilmaisun osa jätetään pois ilman, että merkitys muuttuu
  - "Kallella ei ole koiraa, mutta Arilla on."

9

Luonnollisen kielen ominaisuuksia

- **hyponymia:** abstraktimpaa käsitettä käytetään konkreettisemman sijasta
  - "Mikon auto vietiin romuttamoon. **Kaverin kulkuneuvo** taisi olla huonolla hoidolla."
- **yhdyssanat ja sanaliitot**
  - "kulkuneuvo", "Euroopan unioni"
  - loppuosa ilmoittaa pääluokan ja alkuosa alaluokan: loppuosan tunnistaminen on usein tärkeää

10

Luonnollisen kielen ominaisuuksia

- **etu- ja loppuliitteet**
  - esi- , jälki-, etu-, epä-, -inen, -lainen
  - käyttö ja tärkeys eri kielissä vaihtelee
- **sanojen taipuminen**
- **sanojen johdokset:** sanoista saadaan johtimien avulla uusia sanoja
  - kirja → kirjain, kirjasto, kirjaimisto,…
- **kirjoitusvirheet**

11

Suomen kielen erityisongelmia

- tyypillisiä ongelmia:
  - monitulkintaisuus, erityisesti taivutusmuotohomografia ("hauissa", "puhelin")
  - sanojen ja niiden vartaloiden taipuminen
  - yhdyssanat ja sanaliitot
  - johdokset

12

Suomen kielen erityisongelmia

- taivutusmuotohomografia
  - kahdella perusmuotoisella eri sanalla on joitakin yhteisiä taivutusmuotoja
    - puhelin → puhella, puhelin
    - ilmaisin → ilmaisin (laite), ilmainen, ilmaista
  - pelkkää sananmuodon yhteydestään irrotettua esiintymää tutkimalla ei voi tietää, mikä on perusmuoto
    - on tutkittava sanan kontekstia
  - "Minä en voi syödä voita"
  - "Mänty ja kuusi kestävät ilman saasteita…"

13

Suomen kielen erityisongelmia

- sanojen taipuminen
  - suomenkielessä substantiivi voi taipua n. 2000 eri muodossa, adjektiivi n. 6000 eri muodossa ja verbi n. 12000 eri muodossa
    - mukaan ei ole laskettu johdoksia
  - englanninkielessä voidaan poistaa sanoista kaikki pääteainekset (typistys, stemming)
    - kaikille sananvartaloille yhteinen vartalo
  - suomenkielessä vartalotkin taipuvat usein ja mutkikkaammin (mm. astevaihtelun takia)
    - katto – katon, jalka – jalan, yksi – yhtenä, yö – öitä

14

Suomen kielen erityisongelmia

- yhdyssanat
  - suomenkielessä on paljon yhdyssanoja
  - englanninkielessä usein vastaava asia ilmaistaan sanaliiton avulla
    - istuinvyö – seat belt
    - maitopullo – milk bottle
- johdokset
  - suomenkielessä käytetään paljon johdoksia
  - johdoksen kantasana voidaan tunnistaa ja eri johdokset ryhmitellä yhteen
  - johdokset voivat olla niin vakiintuneita, ettei niitä enää mielletä saman kantasanan johdoksiksi

15

Kielten tyypitys morfologian perusteella

- morfologia:
  - kuvaa sanojen sisäistä rakennetta
  - sana voi jakautua useaan rakenneosaan, morfeemiin
    - talo+ssa
  - morfeemi on pienin rajattavissa oleva osa, jolla on merkitys tai kieliopillinen funktio
  - morfeemi on abstraktio, joka voi esiintyä erilaisina (allo)morfeina
    - esim. vesi, vete, vede, vet, ves ovat saman morfeemin allomorfeja
  - sanoja ei voi aina segmentoida morfeiksi
    - vrt. hoppa + imperfekti = hoppa+de ja brinna + inperfekti = brann

16

Kielten tyypitys morfologian perusteella

- kielet voidaan tyypitellä morfologian suhteen neljään luokkaan
  - isoloivat (isolating) kielet: sanat eivät taivu
  - agglutinoivat (agglutinative) kielet: taivutuspäätteet liitetään "mekaanisesti" vartaloon ja toistensa perään
  - fuusioivat (fusional) kielet: taivutusmuotoja ilmaisevien päätteiden yms. raja hämärä (take vs. took)
  - polysynteettiset (polysynthetic) kielet: paljon taivutuspäätteitä yhdessä sanassa; yksi sana voi vastata lausetta muissa kielissä
- luokittelu ei ole ehdoton: yleensä kielissä on piirteitä useista luokista

17

Kielten tyypitys morfologian perusteella

- luokittelu voidaan tehdä myös tarkastelemalla taipumiseen vaikuttavia piirteitä
  - synteesi-indeksi: kielen sanoissa olevien morfeemien keskimääräinen määrä
    - kuvaa taivutuspäätteiden määrää kielessä
    - isoloivilla kielillä pienin, polysynteettisillä kielillä suurin
  - fuusioindeksi: niiden sanojen suhteellinen osuus, joissa morfeemeja ei voi erottaa toisistaan segmentoimalla
    - fuusioivilla kielillä suurin, agglutinoivilla kielillä pienin
- yo. indeksejä voidaan tarkastella erikseen myös taivutuspäätteiden, johdospäätteiden ja yhdyssanojen kannalta

18

Kielten tyypitys morfologian perusteella

- sanan taivutus ei muuta sanan merkitystä
- johdoksen merkitys voi olla kaukanakin kantasanan merkityksestä
- yhdyssanan merkitys voi olla osasanojensa merkitysten summa tai yhdyssana voi tarkoittaa jotain aivan muuta
  - satakieli, jordgubbe (mansikka)
- kielen johdoksille ja yhdyssanoille voidaan laskea myös semanttinen fuusioindeksi

19

Kielten tyypitys morfologian perusteella

- kielten piirteiden tarkastelua voidaan hyödyntää tiedonhakujärjestelmiä rakennettaessa, mm.
  - millaista morfologista käsittelyä tarvitaan, miten tehokasta se todennäköisesti kullekin kielelle on ja missä ongelmat todennäköisesti tulevat vastaan
  - jos järjestelmässä käytetään useita kieliä, voidaan eri kielillä tehtyjen hakujen tuloksia verrata ja selittää

20

Luonnollisen kielen käsittely hakua varten

- luonnollisen kielen automaattisen käsittelyn vaiheet:
  - sananmuotojen morfologinen analyysi
  - morfologisten moniselitteisyyksien ratkonta
  - syntaktinen analyysi
  - lauseiden kirjaimellisen merkityksen tulkinta
  - lauseen kontekstimerkityksen tulkinta

21

Luonnollisen kielen käsittely hakua varten

- sananmuotojen morfologinen analyysi
  - sanojen palauttaminen perusmuotoon ja tietoihin taivutuspiirteistä
  - esim. "menisi" on "mennä"-verbin konditionaalin preesensin 3. persoonan muoto ("hän men+isi") tai sen kieltomuoto ("hän ei menisi")
- morfologisten moniselitteisyyksien ratkonta eli disambiguiointi
  - lauseessa "Hän menisi sinne" sananmuoto "menisi" on yllämainittu konditionaalin preesensin 3. persoonan muoto (ei vastaava kieltomuoto)

22

Luonnollisen kielen käsittely hakua varten

- morfologisesti analysoitujen lauseiden syntaktinen analyysi eli jäsennys
  - määrittelee kunkin sananmuodon syntaktisen tehtävän lauseessa
  - "Hän menisi sinne.": "hän" on subjekti, "menisi" on predikaatti ja "sinne" on adverbiaali
- lauseiden kirjaimellisen merkityksen tulkinta
  - verbi "mennä" yllä tarkoittaa konkreettista liikettä, jonka myötä elävä olio siirtyy, usein paikasta toiseen
  - verbillä "mennä" on myös muita merkityksiä kuten "sujua"

23

Luonnollisen kielen käsittely hakua varten

- lauseen kontekstimerkityksen tulkinta
  - sana "hän": henkilö, josta on ollut puhetta ja joka ehkä voidaan tarkasti tunnistaa
  - sana "sinne" viitannee paikkaan, joka selviää kontekstin avulla

24

Luonnollisen kielen käsittely hakua varten

- lauseiden ja tekstien merkityksen tulkinta edellyttää yleensä arkitietämystä (common sense knowledge) maailmasta, esineistä ja niiden suhteista
  - vaikea automatisoida
- merkityksen tulkinta ei ehkä edes ole välttämätöntä tiedonhaussa
  - olisi vaikea kuvata kyselyssä, millaisia merkityksiä tuloksen dokumenteissa pitäisi olla
  - merkitykset syntyvät usein myös vasta lukutilanteessa
- sanojen taipumisen hallinta tärkeää
  - sanojen perusmuotoistaminen ja vartaloiden tuottaminen

25

Luonnollisen kielen käsittely hakua varten

- sanojen perusmuotoistaminen
  - verbit → I infinitiivi, nominit → yksikön nominatiivi
  - voidaan myös tunnistaa yhdyssanojen osasanat ja tuottaa johdoksista kantasanan perusmuoto
  - perustuu yksittäisten sananmuotojen analyysiin sanakirjan avulla → homografia ja taivutusmuotohomografia aiheuttavat monitulkintaisuutta
  - menetelmät tuottavat yleensä kaikki tulkintavaihtoehdot, mutta syntaktisella analyysilla voidaan vaihtoehtoja vähentää
  - sanakirja määrää sen, mitä sanoja voidaan tunnistaa

26

Luonnollisen kielen käsittely hakua varten

- taivutusvartaloiden tuottaminen
  - tunnistetaan sanan kaikille taivutusmuodoille yhteinen osa
  - taivutusvartalot saadaan sanojen taipuvista muodoista karsimalla pois kaikki pääteainekset
    - lapsi → lapsi-, lapse-, last-
    - kova → kova-, kovem-, kovi-, kovo-
    - yö → yö-, öi-
  - suosittua englanninkielessä, esim. Porter stemmer
    - vartalot eivät ole välttämättä kieliopillisia

27

Luonnollisen kielen käsittely hakua varten

- suomenkielessä taivutusvartaloiden tuottaminen on monimutkaisempaa
  - pelkkä päätteiden luetteleminen ja karsinta ei toimi
  - esim. –sti –päätteellä saadaan adjektiiveista adverbeja, mutta myös muissa sanoissa on sama pääte (posti, viesti)
  - vartaloiden muodostaminen lähtee perusmuotoisista sanoista, joiden sanaluokka on tunnettu
    - tuotetaan joukko sanan vaihtoehtoisia taivutusvartaloita (mahdollisimman pitkiä)
  - sanaluokkien tunnistamiseksi tarvitaan syntaktista analyysia

28

Käänteistiedostoratkaisut

- sanojen taipuminen ja yhdyssanat voidaan ottaa eri tavoin huomioon käänteistiedostoja rakennettaessa
- Alkula ja Honkela ovat tutkineet käänteistiedostoratkaisujen vaikutusta hakutuloksiin täystäsmäyttävässä ympäristössä
- tarkastellut käänteistiedostoratkaisut:
  - (H1) Sanat ovat taipuneessa muodossaan
  - (H2) Sanat ovat perusmuodossaan
  - (H3) Sanat perusmuodossaan, yhdyssanat ositettuna

29

Käänteistiedostoratkaisut

- käänteistiedostoihin yhdistettiin erilaisia kyselyratkaisuja:
  - (T1) haku hakijan itse katkaisemilla hakuavaimilla (käänteistiedosto H1)
  - (T2) haku automaattisesti tuotetuilla hakuavainten taivutusvartaloilla (kääntdsto H1)
  - (T3) haku automaattisesti tuotetuilla hakuavainten taivutusvartaloilla ja haun jälkeinen seulonta perusmuotoistamalla täsmäävät sananmuodot (kääntdsto H1)
    - varmistetaan, että täsmäävä sananmuoto on hakuavaimen taivutusmuoto

30

Käänteistiedostoratkaisut

- (T4) haku hakuavainten perusmuotojen ja yhdyssanojen alkuosien avulla (kääntdsto H2)
- (T5) haku hakuavainten perusmuotojen ja yhdyssanojen kaikkien osien avulla (kääntdsto H3)

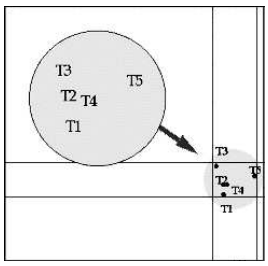
31

Käänteistiedostoratkaisut

- tuloksia
  - suhteellisen saannin kannalta parhaita ympäristöjä olivat T4 ja T5
  - T1: saanti hyvä, mutta tarkkuus huono
  - T3: paras tarkkuus, saanti huonoin
  - erot eivät olleet kovin suuria
    - saanti 80% - 96%
    - tarkkuus 27% - 39%

32

Käänteistiedostoratkaisut



33

Käänteistiedostoratkaisut

- eri ympäristöissä saavutettavien tulosten välillä ei ole suurta eroa → paras vaihtoehto on se, jossa kyselyt on helpointa toteuttaa
- jos sanat on perusmuotoistettu, hakijan ei tarvitse miettiä sanojen taipumista
- kaikkia dokumentin sanoja ei kuitenkaan onnistuta palauttamaan perusmuotoon
- jos yhdyssanoja ei ositeta, kyselyssä on annettava sopivia yhdyssanoja
- sanojen moniselitteisyyden vaikutus pienenee, kun hakusanoja on useita

34

Tässä osassa

- esimerkki täydellisen linkityksen kriteerin käytöstä ryvästämisessä (Saltonin kirjasta)
- luonnollisen kielen käsittely hakua varten
  - luonnollisen kielen piirteitä
    - suomenkielen erityispiirteitä
    - kielten tyypittely morfologisten piirteiden suhteen
  - luonnollisen kielen käsittelymenetelmiä
  - käänteistiedostoratkaisuja

35