ARTIKKELI

Ari Pirkola

Kyselyrakenteiden ja erikoissanakirjan vaikutus sanakirjakäännökseen perustuvassa kieltenvälisessä tiedonhaussa

Kyselyrakenteiden ja erikoissanakirjan vaikutus sanakirjakäännökseen perustuvassa kieltenvälisessä tiedonhaussa [The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval]. Informaatiotutkimus 17 (3): 48-58.

In this study, the effects of query structure and various dictionary-based translation methods on the performance of cross-language information retrieval (CLIR) were tested. Query types studied were concept based, i.e., Boolean queries, and structured and unstructured natural language queries. The structuring of natural language queries was done on the basis of the output of dictionaries. Three translation methods, using a general dictionary and a domain specific (=medical) dictionary, were tested. The document collection was a subset of TREC collection, and as test requests the study used TREC's health related topics. The performance of Finnish test queries, which were automatically translated into English, against English documents was compared to the performance of original English queries against English documents. There was only a slight difference in performance between the original English queries and the best cross-language queries, i.e., concept based queries and structured natural language queries, with medical dictionary and general dictionary translation.

Address: University of Tampere. Department of Information Studies. P.O.Box 607, FIN-33101 Finland. E-mail: pirkola@cc.jyu.fi.

Johdanto

Kun tarve ja mahdollisuudet saada vieraskielistä tietoa ovat tiivistyneen kansainvälisen yhteistyön ja tietoverkkojen kehittymisen myötä tuntuvasti lisääntyneet, on tiedonhaun tutkimuksen piirissä herännyt kiinnostus kieltenvälistä tiedonhakua (cross-language information retrieval, CLIR) kohtaan. Keskeisenä piirteenä kieltenvälisessä tiedonhaussa on, että tiedonhakija esittää kyselyn eri kielellä kuin millä järjestelmän dokumentit on kirjoitettu. Siten suomi-englanti CLIR-järjestelmässä palautteena suomenkieliseen kyselyyn järjestelmä antaa englanninkielisiä dokumentteja. Dokumenttikokoelma voi olla yksikielinen (kuten englanti), jolloin kyselykieliä on yleensä kaksi (kuten suomi ja englanti) tai monikielinen,

jolloin kyselykieliä voi olla useita.

Esittelen tässä artikkelissa kieltenvälistä tiedonhakua koskevan tutkimukseni tuloksia. Johdantona artikkelin tutkimusta käsittelevään osaan kerron kieltenvälisen tiedonhaun ajatelluista hyödyistä, kyselyjen käännösmenetelmistä ja kääntämiseen liittyvistä ongelmista.

Hyödyt, käännösmenetelmät ja ongelmat

Mitä hyötyä kieltenvälisestä tiedonhausta voisi olla? Oard ja Dorr (1996) valaisevat kysymystä mm. seuraavin esimerkein:

-Kun dokumenttikokoelma on monikielinen, on helpompi esittää kysely yhdellä kielellä kuin järjestelmän jokaisella kielellä erikseen.

-Kuvatietokannan kuvat voivat olla erikielisistä lähteistä peräisin, jolloin kuvatekstikieliä on useita. Kun kuvatietokanta on CLIR-perusteinen, hakija voi käyttää yhtä kieltä (äidinkieltään) kuvahaussa.

-Tiedonhakija ei hallitse dokumenttikokoelman kieltä niin hyvin, että pystyisi sujuvasti esittämään kyselyn sillä, mutta ymmärtää sitä kuitenkin riittävän hyvin pystyäkseen hyödyntämään kokoelman dokumentteja.

Tarkastellaan lähemmin viimeistä kohtaa. Useimmat tietävät kokemuksesta, että vieraalla kielellä kirjoittaminen on vaikeampaa kuin vieraskielisen tekstin ymmärtäminen. Aktiivinen ja passiivinen kielen osaaminen ovatkin kaksi aivan eri asiaa. Kyselyn muodostaminen kuuluu aktiivisen kielen osaamisen piiriin. Sen takia vieraskielisen kyselyn muodostaminen saattaa olla vaikeaa, vaikka kysyjä hallitsisikin vieraan kielen niin hyvin, että hän ymmärtää asiasisällön lukemastaan.

Tulevaisuuden rutiinia saattaa olla dokumenttien automaattinen kääntäminen CLIR-järjestelmän yhteydessä. Tällöin järjestelmä kääntää haetut vieraskieliset dokumentit tiedonhakijan äidinkielelle. Hakijan ei tarvitse osata dokumenttien kieltä lainkaan, eipä edes tietää sitä, mikä kieli on kyseessä, mutta hän voi kuitenkin automaattisen käännöksen ansiosta käyttää hyväksi löydettyjä dokumentteja.

Tiedonhaku kielten välillä vaatii sen, että joko kyselyt käännetään dokumenttien kielelle tai dokumentit käännetään kyselyn kielelle. Edellinen on helpompaa ja tietokoneen kapasiteettia vähemmän rasittavaa, minkä johdosta alan tutkimus lähtee kyselyjen kääntämisestä. Perusmenetelmät kyselyn kääntämisessä ovat automaattinen konekäännös, korpukseen perustuvat käännösmenetelmät ja sanakirjakäännös.

Nykyisten automaattisten konekäännösjärjestelmien tuottamien käännösten taso ei useinkaan ole kovin korkeatasoinen (Hull & Grefenstette, 1996; Oard & Dorr, 1996; Yamabana et al., 1996). Rajatuilla aihealueilla järjestelmä voi kuitenkin tarjota riittävän spesifistä kääntämisessä tarvittavaa tietoa, jolloin on mahdollista saada korkeatasoisia käännöksiä. Perusongelmana konekäännöksen soveltamisessa kyselyjen kääntämiseen on kuitenkin se, että kyselyt ovat yleensä lyhyitä, usein muutaman sanan pituisia ja vailla kieliopillisesti oikeaa syntaktista rakennetta. Tällöin on mahdoton soveltaa syntaktista analyysia, johon konekäännös suureksi osaksi perustuu.

Korpukseen perustuvissa menetelmissä kyselyt käännetään ja niitä laajennetaan kaksi- tai monikielisten sanalistojen avulla, jotka saadaan paralleeleista korpuksista (parallel corpora) tai vastindokumentteja sisältävistä kokelmista (comparable document collections). Paralleeleissa korpuksissa jokainen kokoelman dokumentti on käännetty kaikille (kummallekin) kokoelman kielille. Vastindokumenttikokoelmat käsittävät samaan aihepiiriin liittyviä erikielisiä dokumentteja. Kääntäminen perustuu siihen, että identtiset ja sama-aiheiset erikieliset dokumentit sisältävät vastinsanoja. Siis jos suomenkielisessa dokumentissa esiintyy sana *talo*, niin vastaavassa englanninkielisessa dokumentissa on sana *house*. Vastinsanat poimitaan vastindokumenteista tilastollisin menetelmin.

Korpukseen perustuvat kieltenväliset kyselyt ovat antaneet vaihtelevia tuloksia (Davis, 1997; Davis & Dunning, 1996; Dumais et al., 1996; Sheridan & Ballerini, 1996; Sheridan et al., 1997). Erittäin hyviäkin tuloksia on saatu, sillä Sheridanin (et al., 1997) tutkimuksessa kieltenvälisten kyselyjen tehokkuus oli lähes yhtä hyvä kuin vastaavien yksikielisten kyselyjen tehokkuus.

Sanakirjakäännöksessä käytetään elektronisia kaupallisia kaksikielisiä sanakirjoja, jotka on muutettu CLIR-ympäristöön sopiviksi karsimalla niistä ylimääräinen tekstiaines pois. On myös olemassa kaksi- ja monikielisiä tesauruksia, jotka on varta vasten laadittu kieltenvälistä tiedonhakua varten (Gilarranz et al, 1997). Seuraavaksi tarkastellaan varsinaista sanakirjakäännöstä.

Elektronisen sanakirjan avulla tapahtuvassa kääntämisessä lähdekielen sana korvataan kaikilla sanakirjaan sisältyvillä kohdekielen vastinsanoilla, jotka kaikki otetaan mukaan lopulliseen kyselyyn (CLIR-kyselyyn). Tutkimukset ovat osoittaneet, että sanakirjan avulla muodostetut CLIR-kyselyt ovat tehottomia. Useat tutkijat ovat saaneet tuloksia, joiden mukaan yksinkertaiseen sanakirjakäännökseen perustuvien CLIR-kyselyjen tarkkuus on vain noin puolet vastaavien yksikielisten kyselyjen tarkkuudesta (Ballesteros & Croft, 1996; Davis & Dunning, 1996; Hull & Grefenstette, 1996).

Sanakirjakäännökseen liittyvät perusongelmat ovat (1) sanaliittojen tunnistaminen, (2) käännös-polysemia ja (3) sanakirjojen kattavuus. Luonnollisen kielen hauissa tiedonhakija voi esittää kyselyn luonnollisilla lauseilla. Jollei järjestelmään liity sanaliittojen tunnistamismenetelmää (ks. Ballesteros & Croft, 1997), sanakirja kääntää sanaliittojen sijasta sanaliittojen komponenttisanat. Tällöin sanaliittojen merkitys usein häviää. Tällä on tarkkuutta heikentävä vaikutus. Suomen kielessä sanaliitot eivät kuitenkaan ole kovin iso ongelma, koska moniosaiset sanat ovat suomessa usein muodoltaan yhdyssanoja.

Käännöspolysemiassa on kyse siitä, että sanojen merkitysten määrä lisääntyy, kun lähdekielen sana korvataan kaikilla sanakirjaan sisältyvillä kohdekielen vastinsanoilla. Sen vuoksi sanojen ja sanojen merkitysten määrä on CLIR-kyselyssä usein suurempi kuin lähdekielisessä kyselyssä. Kattavuusongelma koskee varsinkin erikoisalojen termejä, joita yleissanakirjoissa on niukasti. Kuitenkin monissa hakupyynnöissä juuri erikoistermit ovat keskeisiä.

Tutkimusongelmat

Tutkimuksessa tartuttiin käännöspolysemiaongelmaan ja sanakirjojen kattavuusongelmaan käyttämällä kääntämisessä sanakirjakombinaatiota, joka koostui yleissanakirjasta ja erikoisalan, so. lääketieteen, sanakirjasta. Aihealue rajattiin lääketieteeseen ja terveyteen valitsemalla testikysymyksiksi TREC-konferenssin (ks. kohta "tutkimusympäristö") hakutehtäväkokoelmasta lääketiedettä ja terveyttä käsittelevät hakutehtävät. Suomesta englantiin käännettyjen CLIR-kyselyjen tehokkuutta verrattiin alkuperäisten englanninkielisten kyselyjen tehokkuuteen.

Koska aihealueena oli lääketiede ja terveys, lääketieteen sanakirjan odotettiin disambiguoivan sanojen merkityksiä, ts. sen odotettiin antavan vähemmän ja merkitykseltään yksiselitteisimpiä vastinsanoja kuin yleissanakirjan. Kattavuusongelman käytännön merkityksen oletettiin olevan siinä, että yleissanakirjat eivät sisällä erikoisalojen termejä, jotka kuitenkin monissa hakupyynnöissä ovat keskeisiä. Siten osan keskeisistä hakuavaimista oletettiin esiintyvän vain lääketieteen sanakirjassa.

Polysemiaongelmaan tartuttiin myös kyselyjen strukturoinnin kautta. Kyselyjen perustyyppejä oli kaksi: (1) käsiteanalyysiin perustuvat rakenteiset kyselyt (Boolen kyselyt) ja (2) luonnollisen kielen kyselyt. Jälkimmäinen tyyppi jaettiin kahteen alatyyppiin, rakenteettomiin ja rakenteisiin kyselyihin. Aikaisemman tutkimustiedon perusteella oli odotettavissa, että rakenteiset CLIR-kyselyt ovat tehokkaampia kuin rakenteettomat (Hull, 1997). Mainitussa tutkimuksessa kyselyjen strukturointi perustui käsiteanalyysiin.

Tutkimusympäristö

Tutkimusympäristö muodostui seuraavista osista:

-TREC:n dokumentit, relevanssiarviot dokumenteille sekä lääketiedettä ja terveyttä käsittelevät hakutehtävät
-Morfologinen analyysiohjelma TWOL suomen kielelle
-Elektroniset suomi - englanti yleissanakirja ja suomi - englanti lääketieteen sanakirja
-INQUERY[1] tiedonhakujärjestelmä
-Morfologinen analyysiohjelma Kstem englannin kielelle

TREC (Harman, 1996) on kansainvälinen tiedonhaun tutkimuksen foorumi, jossa tutkijat voivat hyödyntää yhteistä tutkimusaineistoa. Tämä käsittää laajan dokumenttikokoelman, useita satoja hakutehtäviä sekä relevanssiarviot dokumenteille.

Testikysymyksiä oli kaikkiaan 34. Tietokannan dokumenttikokoelma sisälsi 514825 dokumenttia, ja tietokannan perustiedoston koko oli 1,46 GB. Kokoelma koostui pääasiassa uutisaineistosta sisältäen *AP Newswire, Federal Register ja DOE abstracts* dokumentteja.

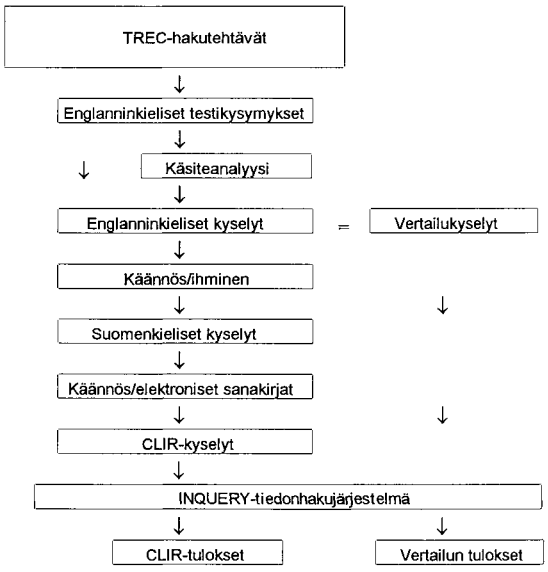
TWOL on morfologisen analyysin suorittava ohjelma, joka tuottaa perusmuotoisia sanoja sanojen taivutusmuodoista sekä pilkkoo yhdyssanat osiinsa. Luonnollisten lauseiden kyselyt sisälsivät taivutusmuotoisia sanoja, jotka ohjelman avulla muutettiin perusmuotoisiksi. Tämä oli tarpeen, koska sanakirjojen hakusanat ovat perusmuotoisia. Yhdyssanat pilkottiin, koska ne esiintyvät sanakirjoissa toisinaan vain osinaan. Sekä yhdyssanat että niiden komponenttisanat käännettiin.

Tutkimuksessa käytetyistä Kielikone Oy:n tuottamista elektronisista sanakirjoista yleissanakirja sisältää 165 000 hakusanaa ja lääketieteen sanakirja 67 000 hakusanaa. Sanakirjojen kaupalliset versiot konvertoitiin CLIR-versioiksi Tampereen yliopiston Informaatiotutkimuksen laitoksen Tiedonhaun laboratoriossa kehitetyllä suodatinohjelmalla. Ohjelma poisti sanakirjoista kaiken muun tekstiaineksen paitsi varsinaiset sanakirjasanat. CLIR-versiot sisälsivät jonkin verran virheitä, minkä johdosta muutamalle suomen kielen sanalle saatiin virheelliset käännökset. Kuitenkin vain alle 1 % käännöksistä oli virheellisiä.

INQUERY on probabilistinen tiedonhakujärjestelmä, joka esittää haun tuloksena saadut dokumentit todennäköisen relevanssin mukaisessa järjestyksessä (Broglio et al., 1994). Kyselyt voidaan esittää luonnollisina lauseina tai ne voidaan formuloida rakenteisiksi järjestelmään sisältyvien monien erilaisten operaattoreiden avulla. Järjestelmään kuuluva morfologinen analyysiohjelma Kstem muuttaa englannin kielen taivutusmuotoiset sanat perusmuotoisiksi.

Tutkimuksen kulku

Yleiskuva tutkimuksen kulusta on kuvassa 1. Kuvan "käsiteanalyysi" -vaihe koskee vain käsiteanalyysiin perustuvia kyselyjä. Muuten tutkimusmenetelmät ja -prosessit olivat samat käsiteanalyysiin perustuvissa ja



Kuva 1. Tutkimuksen kulku

luonnollisen kielen kyselyissä. Tutkimuksessa käytettiin CLIR-tutkimuksissa yleisesti käytettyä kyselyvertailut mahdollistavaa menetelmää, jossa alkuperäiset kyselyt (englanti) käännetään elektronisten sanakirjojen lähdekielelle (suomi) ihmisen toimesta. Lähdekieliset kyselyt käännetään sanakirjojen avulla tietokannan dokumenttien kielelle (englanti) CLIR-kyselyiksi. Menetelmän avulla CLIR-kyselyjen tehokkuutta voidaan verrata alkuperäisten kyselyjen tehokkuuteen.

Kyselyjen muodostaminen ja kääntäminen

Testikysymykset, jotka koostuivat 1-2 lauseesta, muodostettiin TREC-hakutehtävien *title-, description-* ja *narrative*-kenttien perusteella. Kyselyjen perustyyppejä oli kaksi. Testikysymykset itsessään olivat ensimmäinen tyyppi. Tästä tyypistä käytetään merkintää LL (luonnollinen lause). Toinen perustyyppi, käsiteperusteiset kyselyt, muodostettiin testikysymysten perusteella poimimalla niistä tärkeimmät sanat ja sanaliitot hakuavaimiksi ja tekemällä käsiteanalyysi. Tämä tyyppi lyhennetään KA. Käsiteanalyysissa määritettiin se, mitkä valituista hakuavaimista edustivat samaa rajaavaa käsitettä hakupyynnössä. Analyysin perusteella muodostettiin käsitteelliset hakusuunnitelmat, jotka rakentuivat hakupyynnön rajaavia käsitteitä edustavista hakuavainryhmistä eli faseteista. Muutamassa testikysymyksessä oli useampi kuin yksi (2-5) näkökulma käsiteltävään aiheeseen. Jokaisesta näkökulmasta tehtiin käsitteellinen hakusuunnitelma.

Artikkelin kirjoittaja käänsi englanninkieliset LL- ja KA-kyselyt suomeksi käyttäen apunaan painettuja sanakirjoja. Elektronisia sanakirjoja ei tässä vaiheessa käytetty. Jokaiselle englannin kielen ilmaisulle annettiin tarkka suomenkielinen vastine. Englanninkieliset LL- ja KA-kyselyt, joiden perusteella suomenkieliset kyselyt muodostettiin, toimivat myös vertailukyselyinä (baselinekyselyinä) CLIR-kyselyille (ks. kuva 1).

LL-kyselyistä muodostettiin kaksi alatyyppiä, rakenteettomat ja rakenteiset LL-kyselyt. Jälkimmäisessä tyypissä sanat fasetoitiin sanakirjojen antamien tulostietueiden perusteella; ne englanninkieliset sanat, jotka vastasivat samaa suomenkielisestä sanaa, ryhmitettiin samaan fasettiin. Rakenteettomat ja rakenteiset LL-kyselyt ovat vertailukelpoisia keskenään, koska ne muodostettiin samojen suomenkielisten kyselyjen perusteella. Myös niiden vertailukyselyt ovat samat. LL- ja KA -kyselyt sitä vastoin eivät ole vertailukelpoisia, koska käytetyt operaattorit ja hakuavaimet eivät ole samat. Lisäksi KA-kyselyjen muodostamisessa on mukana inhimillinen komponentti. Eri ihmiset saattavat näet tunnistaa eri tavalla hakupyyntöjen käsitteet.

Kyselyissä käytetyt operaattorit olivat seuraavat:

-Rakenteettomat LL-kyselyt: *sum, uw3*
-Rakenteiset LL-kyselyt: *sum, syn, uw3*
-Käsiteperusteiset kyselyt: *and, syn, or, uw3*

Tutkimuksessa käytetty tiedonhakujärjestelmä INQUERY esittää haetut dokumentit todennäköisen relevanssin mukaisessa järjestyksessä. Järjestelmän dokumentille laskemaan painoarvoon vaikuttaa kyselyn operaattorit. And-operaattori vaikuttaa siten, että mitä useampi operaattorilla sidottu fasetti (= hakuavain) dokumentissa on, sitä korkeamman painoarvon dokumentti saa. Or-operaattoria käytettiin niissä tapauksissa, joissa testikysymyksillä oli niihin sisältyvien eri näkökulmien johdosta useita käsitteellisiä hakusuunnitelmia. Eri osasuunnitelmia vastaavat hakulausekkeet sidottiin or-operaattorilla toisiinsa yhdeksi kyselyksi. Ainakin yhden or-operaattoriin sisältyvän hakuavaimen on oltava dokumentissa, jotta dokumentille laskettaisiin sen perusteella painoarvo. Sum-operaattorin yhdistämillä hakuavaimilla on yhtäläinen vaikutus hakutuloksiin. Syn-operaattoria sovellettiin samaan fasettiin kuuluviin hakuavaimiin. Järjestelmä pitää sillä sidottuja hakuavaimia saman hakuavaimen esiintyminä. Uwn (unordered window n) on läheisyysoperaattori, ja tutkimuksessa sitä

sovellettiin sanaliittoihin. N-arvoksi valittiin 3, ts. uwn-operaation hakuavaimet rajattiin dokumenteissa kolmen sanavälin etäisyydelle toisistaan.

Käännösmenetelmät olivat:

ysk-käännös: suomenkielisten kyselyjen hakuavaimet käännettiin yleissanakirjan avulla.

esk –>ysk -käännös: suomenkielisten kyselyjen hakuavaimet käännettiin ensiksi lääketieteen sanakirjan ja sitten yleissanakirjan avulla. Yleissanakirjaa sovellettiin vain jos käännettävä ilmaisu ei ollut hakusanana lääketieteen sanakirjassa.

esk ja ysk -käännös: suomenkielisten kyselyjen hakuavaimet käännettiin sekä lääketieteen että yleissanakirjan avulla. Niissä tapauksissa, jolloin molemmat sanakirjat antoivat käännettävälle ilmaisulle saman vastinsanan, vain yksi vastinsanan esiintymä otettiin CLIR-kyselyyn.

Jollei sana tai sanaliitto ollut hakusanana sanakirjoissa, se meni muuttumattomana CLIR-kyselyyn. Tällaiset ilmaisut olivat englanninkielisiä erisnimiä, akronyymejaja suomenkielisiä sanakirjoissa esiintymättömiä sanoja.

Tulokset

Kyselyjen tehokkuus evaluoitiin (1) tarkkuutena 10%-saantitasolla, (2) keskimääräisenä tarkkuutena 10%-100% saantitasoilla ja (3) tarkkuus-saanti -kuvaajina.

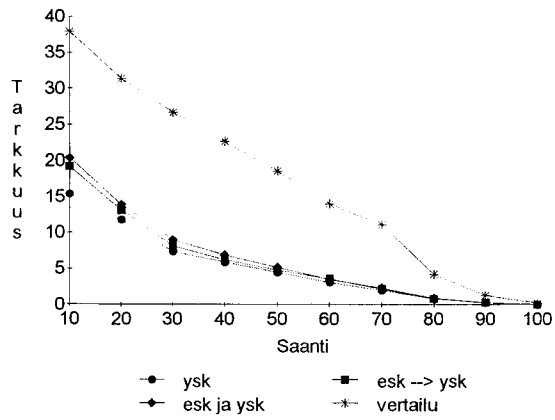
Taulukossa 1 ja kuvissa 2-3 (kuva 3 sivulla 54) on esitetty LL-kyselyjen tulokset. Rakenteettomat ja rakenteiset kyselyt ovat vertailukelpoisia keskenään, koska ne on muodostettu samojen suomenkielisten kyselyjen perusteella. Erona on se, että rakenteisissa kyselyissä on fasettirakenne, mutta rakenteettomissa kyselyissä sitä ei ole.

Rakenteettomien CLIR-kyselyjen tarkkuus on huomattavasti heikompi kuin vertailukyselyjen (taulukko 1 ja kuva 2). 10%-saantitasolla vertailukyselyjen tarkkuus on 37.9%, mutta *ysk*-kyselyjen vain 15.4%. Erikoisanakirjalla on positiivinen vaikutus, sillä *ysk*-kyselyjen tarkkuus on heikompi kuin *esk–> ysk*- ja *esk ja ysk* -kyselyjen. Näidenkin tehokkuus on kuitenkin selvästi alhaisempi kuin vertailukyselyjen.

CLIR-kyselyjen strukturointi sanakirjojen antamien tulostietueiden perusteella parantaa tarkkuutta merkittävästi (taulukko 1 ja kuva 3). Käännösmenetelmistä paras on *esk ja ysk* -menetelmä. *Esk ja ysk* -kyselyjen keskimääräinen tarkkuus on 12.9% ja tarkkuus 10%-saantitasolla 35.9%. Vastaavasti

Kyselytyyppi/Käännösmenetelmä	10%-saanti Ta	Keskim. Ta
Rakenteiset		
YSK	30,9	10,5
ESK --> YSK	30,4	11,3
ESK ja YSK	35,9	12,9
Rakenteettomat		
YSK	15,4	5,1
ESK --> YSK	19,2	5,8
ESK ja YSK	20,4	6,3
Rakenteiset ja rakenteettomat, vertailu	37,9	16,8

Taulukko 1. LL-kyselyjen tehokkuus



Kuva 2. Rakenteettomien LL-kyselyjen tarkkuus–saanti -kuvaajat

vertailukyselyjen keskimääräinen tarkkuus on 16.8% ja tarkkuus 10%-saantitasolla 37.9%. 10%-saantitasolla *esk ja ysk* -kyselyt ovat siten tehokkuudeltaan lähes vertailukyselyjen veroisia. Myös *ysk* ja *esk–>ysk* -kyselyjen rakenteiset versiot antavat huomattavasti paremmat tulokset kuin rakenteettomat versiot.

Myös käsiteperusteisissa CLIR-kyselyissä paras käännösmenetelmä on *esk ja ysk* -menetelmä (taulukko 2, kuva 4 sivulla 55). *Esk ja ysk* -kyselyjen keskimääräinen tarkkuus on 15.5% ja tarkkuus 10%-saantitasolla 32.0%. Vastaavat luvut vertailukyselyillä ovat 16.7% ja 33.6%.

Johtopäätökset

Tutkimuksen tulokset voidaan tiivistää seuraavasti:

1. Sanakirjakäännökseen perustuva kieltenvälinen tiedonhaku voi olla yhtä tehokasta kuin perinteinen yksikielinen tiedonhaku.

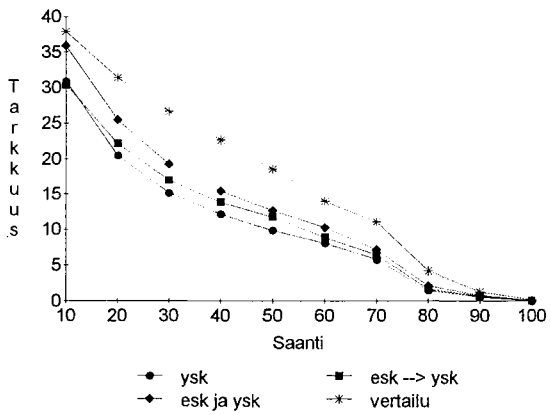
2. Kieltenvälisessä tiedonhaussa kyselyjen strukturointi on ensiarvoisen tärkeää.

3. Tutkimustulokset viittaavat siihen, että sanakirjaperusteinen mekaaninen strukturointi on yhtä tehokasta kuin käsiteanalyysiin perustuva strukturointi. Käsiteperusteisten kyselyjen tehokkuus riippuu kuitenkin kyselyjen tyhjentävyydestä ja kattavuudesta (hakuavainten lukumäärästä fasettia kohden). Näitä tekijöitä tutkimuksessa ei tarkasteltu.

4. Erikoissanakirjalla on positiivinen vaikutus sanakirjakäännökseen perustuvassa kieltenvälisessä tiedonhaussa.

5. Kun dokumenttikokoelma sisältää uutisaineistoa, erikoissanakirja- ja yleissanakirjakäännös on selvästi parempi käännösmenetelmä kuin menetelmä, jossa ensiksi sovelletaan erikoissanakirjaa ja sitten yleissanakirjaa, tai pelkkä yleissanakirjakäännös.

Sanakirjakäännökseen perustuvien CLIR-kyselyjen tehokkuuden on todettu olevan selvästi heikompi kuin yksikielisten kyselyjen. Hull ja Grefenstette (1996) tutkivat ranskasta englantiin käännettyjä CLIR-kyselyjä. Nämä käännettiin automaattisesti tuotetun yksittäisiä sanoja sisältävän sanakirjan avulla. CLIR-kyselyjen tarkkuus oli vain 60% alkuperäisten englanninkielisten kyselyjen tarkkuudesta. Sama sanakirjan avulla muodostettujen CLIR-kyselyjen tehottomuus on havaittu muissakin tutkimuksissa (Ballesteros & Croft, 1996; Davis, 1997; Davis &



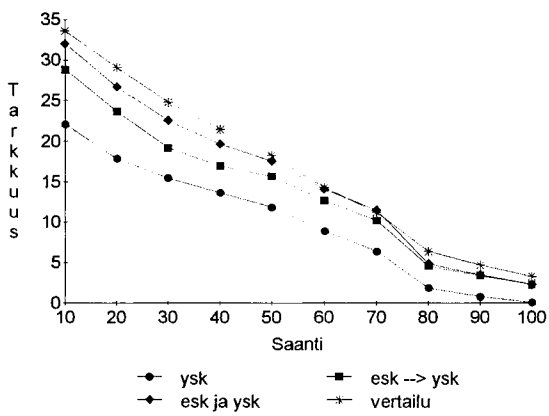
Kuva 3. Rakenteisten LL-kyselyjen tarkkuus–saanti -kuvaajat

Käännösmenetelmä	10%-saanti Ta	Keskim. Ta
YSK	22,1	9,9
ESK –> YSK	28,8	13,8
ESK ja YSK	32,0	15,5
Vertailu	33,6	16,7

Taulukko 2. Käsiteperusteisten kysely jen tehokkuus

Dunning, 1996). Davisin (1997) mukaan sanaluokkaan perustuva disambiguointi eli käännösvastineiden poimiminen sanakirjasta siten, että kohdekielen sanan sanaluokka vastaa lähdekielen sanan sanaluokkaa, parantaa CLIR-kyselyjen tarkkuutta selvästi. Vielä tästä parempi menetelmä on sanakirjakäännös sanaluokkaan ja korpukseen perustuvan disambiguoinnin kanssa. Davisin tutkimuksessa keskimääräiset tarkkuusarvot olivat seuraavat: pelkkä sanakirjakäännös, 0.14; sanakirjakäännös täydennettynä sanaluokka-disambiguoinnilla, 0.19; sanakirjakäännös täydennettynä sanaluokka-disambiguoinnilla sekä korpukseen perustuvalla disambiguoinnilla, 0.21; yksikieliset vertailukyselyt, 0.29. Hullin (1997) mukaan sanakirjalla käännetyt rakenteiset (Boolen) kyselyt antavat paremman tuloksen (keskimääräinen tarkkuus 0.28) kuin sanakirjalla käännetyt vektorimalliin perustuvat kyselyt (keskimääräinen tarkkuus 0.20).

Tämän tutkimuksen tulokset osoittavat, että sanakirjakäännökseen perustuva kieltenvälinen tiedonhaku on lähes yhtä tehokasta kuin perinteinen yksikielinen tiedonhaku, jos kääntämisessä käytetään yleissanakirjan lisäksi erikoissanakirjaa ja jos kyselyt strukturoidaan sanakirjojen antamien tulosten tai käsiteanalyysin perusteella. Koska tutkimuksessa käytettyjen elektronisten sanakirjojen kaupalliset versiot konvertoitiin automaattisesti CLIR-versioiksi eikä CLIR-versioita tai käännöksiä muutettu manuaalisesti, tutkimuksen kyselyjen tehokkuustaso on mahdollista saavuttaa käytännössä operationaalisessa CLIR-järjestelmässä.



Kuva 4. Käsiteperusteisten kyselyjen tarkkuus–saanti -kuvaajat

Millä tavalla strukturointi sitten vaikuttaa? Tärkein tekijä näyttää olevan se, että strukturoinnin ansiosta kyselyn tärkeiden hakuavainten suhteellinen paino lisääntyy. Usein ne hakuavaimet, joilla on yksi tai kaksi käännösvastinetta, ovat kyselyn tärkeimpiä avaimia, ja toisaalta ne hakuavaimet, joilla on useita käännösvastineita, ovat vähemmän merkityksellisiä avaimia (Hull, 1997). Siten rakenteettomassa CLIR-kyselyssä toisarvoiset hakuavaimet ja epärelevantit käännösvastineet dominoivat aiheuttaen sen, että tärkeiden hakuavainten suhteellinen paino ja vaikutus hakutulokseen jää vähäiseksi. Seurauksena on alhainen tarkkuus. Strukturoinnin jälkeen kyselyn hakuavaimet eivät enää ole tasavertaisia, vaan nyt tasavertaisuus vallitsee fasettien välillä. Tällä tavalla tärkeät hakuavaimet saavat suhteellisesti enemmän painoa. Strukturoinnin vaikutus kyselyn rakenteeseen on demonstroitu artikkelin liitteessä.

Käsiteperusteisia ja luonnollisen kielen kyselyjä ei suoraan voi verrata keskenään. KA- ja LL-kyselyjen suhteellista tehokkuutta voi kuitenkin tarkastella vertailukyselyjen kautta. Suhteessa vertailukyselyihin käsiteperusteiset kyselyt ovat yhtä tehokkaita kuin luonnollisella kielellä esitetyt kyselyt. Kun tarkastellaan *esk ja ysk* -käännösmentelmää ja 10%-saantitasoa, KA-kyselyjen suhteellinen tehokkuus on 32.0/33.6 = 95.2. LL-kyselyissä suhde on 35.9/37.9 = 94.7. On kuitenkin huomattava, että käsiteperusteisten kyselyjen tehokkuus riippuu niiden kattavuudesta ja tyhjentävyydestä. Tutkimuksen käsiteperusteiset kyselyt olivat suppeita, sillä useissa tapauksissa rajaava käsite oli ilmaistu vain yhdellä hakuavaimella, sekä tyhjentäviä, sillä hakupyynnön kaikki rajaavat käsitteet olivat mukana kyselyissä. Lisäksi on syytä huomata, että käsiteperusteisissa kyselyissä oli mukana inhimillinen komponentti, sillä käsiteanalyysin tulos riippuu analyysin tekijästä.

Lääketieteen sanakirjan aiheuttama CLIR-

kyselyjen tehokkuuden paraneminen johtui ensiksikin siitä, että sanakirja sisälsi hakuavaimia, joita ei ollut yleissanakirjassa. Erikoissanakirjoille tyypillinen piirre on, että niissä on runsaasti spesifejä erikoistermejä, kun taas yleissanakirjoissa niitä on vähän. Moniin hakupyyntöihin sisältyy erikoistermejä ja usein ne ovat hakupyynnön keskeisiä ilmaisuja. Yksi tutkimuksen testikysymys oli: "Mitä osteoporoosin ehkäisyyn ja sairauden aiheuttamien seurausten lieventämiseen liittyviä tutkimuksia on meneillään?". Tärkein ilmaisu on selvästikin termi *osteoporoosi*. Termi esiintyy tutkimuksessa käytetyssä lääketieteen sanakirjassa, mutta sitä ei ole yleissanakirjassa. Myös joissakin muissa tutkimuksen testikysymyksissä oli lääketieteellisiä erikoistermejä, jotka esiintyivät vain lääketieteen sanakirjassa. Erikoistermejä sisältävät kyselyt antoivatkin huonon tuloksen silloin, kun kääntäminen suoritettiin pelkän yleissanakirjan avulla.

Toiseksi, lääketieteen sanakirjan aikaansaama CLIR-kyselyjen tehokkuuden paraneminen johtui sanakirjan käännöspolysemiaa vaimentavasta vaikutuksesta. Yleissanakirjoille on tunnusomaista, että ne antavat käännettävälle sanalle useita käännösvastineita. Erikoissanakirjat sitä vastoin antavat usein yhden tai kaksi vastinsanaa. Yleissanakirjojen sanoilla on usein monia merkityksiä, kun sen sijaan erikoissanakirjan sanoilla merkityksiä on vähän, usein yksi spesifi merkitys. Näistä syistä käännöspolysemia ei erikoissanakirjakäännöksissä ole yhtä suuri ongelma kuin yleissanakirjakäännöksissä. Esimerkiksi tutkimuksen yleissanakirja antoi sanalle *leikkaus* seitsemän käännösvastinetta: *cut, cutting, clipping, operation, editing, section,* and *retrenchment*. Yhteenlaskettujen sananmerkitysten määrä on huomattavasti enemmän kuin 7. Lääketieteen sanakirja antoi vain kaksi vastinetta: *operation* and *surgery*. Sananmerkityksiä on enemmän kuin kaksi, mutta kuitenkin selvästi vähemmän kuin edellisessä tilanteessa.

Joissakin tapauksissa lääketieteen sanakirjalla oli haitallisia vaikutuksia. Ensiksi, se antoi kahdelle sanalle virheelliset käännösvastineet, ja toiseksi, jotkin sen antamista käännöksistä olivat ilmaisuja, joita ei arkielämässä juurikaan käytetä. Varsinkin jälkimmäinen tekijä on syynä siihen, miksi *esk ja ysk* -kyselyt antoivat paremmat tulokset kuin *esk–>ysk* -kyselyt. Jälkimmäisissä kyselyissähän yleissanakirjakäännöstä ei suoritettu, jos lääketieteen sanakirja käänsi ilmaisun. Käännösmenetelmän tehokkuus riippuu kuitenkin tietokannan dokumenttien tyypistä. Tutkimuskokoelman dokumentit olivat valtaosaltaan uutisdokumentteja. Tämän tyyppinen teksti sisältää sekä tieteellisiä että arkipäivän ilmaisuja. Tällaiseen tekstityyppiin erikois- *ja* yleissanakirjakäännös soveltuu hyvin. *Esk–>ysk* -käännösmenetelmä sopii todennäköisesti hyvin tieteellisiin teksteihin, sillä menetelmässä erikoissanakirjaa painotetaan.

Tutkimuksen testikysymyksiksi otettiin vain ne TREC:n hakutehtävät, jotka käsittelivät lääketiedettä ja terveyttä. Tämä ei kuitenkaan merkitse sitä, että nyt kuvattu järjestelmä pystyisi käsittelemään vain yhden erikoisalan tekstejä. Järjestelmää voidaan laajentaa niin, että se kykenee prosessoimaan monien alojen kyselyjä ja dokumentteja. Itse asiassa siitä on mahdollista tehdä niin laaja, että se kattaa elämän koko tekstikirjon. Järjestelmään voidaan liittää yleissanakirja sekä useiden alojen erikoissanakirjoja. Kyselyjen erikoistermit voidaan osoittaa automaattisesti oikeisiin sanakirjoihin, koska eri alojen terminologiat poikkeavat toisistaan. Aihealuepolysemia, termin esiintyminen usealla alalla eri merkityksissä, lienee harvinaista. Vaihtoehtoisena menetelmänä tiedonhakija voisi informoida järjestelmää kyselynsä aihealueesta.

Tutkimustulokset osoittavat, että sanakirjakäännökseen perustuva kieltenvälinen tiedonhaku on lähes yhtä tehokasta kuin perinteinen tiedonhaku, jos kääntämisessä käytetään erikois- ja yleissanakirjaa ja jos kyselyt strukturoidaan sanakirjojen antamien tulosten tai käsiteanalyysin perusteella. Näillä menetelmillä käännöspolysemiaongelma sekä sanakirjojen kattavuusongelma voidaan ratkaista hyvällä menestyksellä.

Hyväksytty julkaistavaksi 16.3.1998.

Kirjallisuus

Ballesteros, L. & Croft, W.B.(1997). Phrasal Translation and Query Expansion techniques for Cross-Language Information Retrieval. Working Notes of AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, Stanford, CA, s.1-8.

Ballesteros, L. & Croft, W.B. (1996). Dictionary-based methods for cross-lingual information retrieval. Proceedings of the 7th International DEXA Conference on Database and Expert Systems Applications, s. 791-801.

Broglio, J., Callan, J. & Croft, W.B. (1994). Inquery system overview. Proceedings of the TIPSTER Text Program (Phase I), s. 47-67.

Davis, M. (1997). New experiments in cross-language text retrieval at NMSU's Computing Research Lab. Proceedings of the Fifth Text REtrieval Conference (TREC-5). Ed. D.K. Harman. Gaithesburg, MD.

Davis, M. & Dunning, T. (1996). A TREC-evaluation of query translation methods for multi-lingual text retrieval. Proceedings of the Fourth Text REtrieval Conference (TREC-4). Ed. D.K. Harman. Gaithesburg, MD, s. 483-497.

Dumais, S.T., Landauer, T.K. & Littman, M.L. (1996). Automatic cross-linguistic information retrieval using latent semantic indexing. Working Notes of the Workshop on Cross-Linguistic Information Retrieval. Eds. G. Grefenstette & A. Smeaton & P. Sheridan, ACM SIGIR, Zürich, s. 16-23.

Gilarranz, J., Gonzalo, J. & Verdejo, F. (1997). An Approach to Conceptual Text Retrieval Using the EuroWordNet Multilingual Semantic Database. Working Notes of AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, Stanford, CA, s. 51-57.

Harman, D. (1996). Overview of the Fourth Text REtrieval Conference (TREC-4). The Fourth Text REtrieval Conference (TREC-4). Ed. D.K. Harman. s. 1- 23.

Hull, D. (1997). Using Structured Queries for Disambiguation in Cross-Language Information Retrieval. Working Notes of AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, Stanford, CA, s. 73-81.

Hull, D. & Grefenstette, G. (1996). Querying across languages. A dictionary-based approach to multilingual information retrieval. Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zürich, s. 49-57.

Oard, D. & Dorr, B. (1996). A Survey of Multilingual Text Retrieval. Technical Report UMIACS-TR-96-19, University of Maryland, Institute for Advanced Computer Studies.

Sheridan, P., Braschler, M. & Schäuble, P. (1997). Cross-Language Information Retrieval in a Multilingual Legal Domain. Research and Advanced Technology for Digital Libraries. Proceedings of the First European Conference, ECDL '97, Pisa, Italy, 1-3 September. Eds. C. Peters & C. Thanos. Lecture Notes in Computer Science, Vol. 1324, s. 253-268.

Sheridan, P. & Ballerini, J. (1996). Experiments in Multilingual Information Retrieval using SPIDER system. Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zürich, s. 58-65.

Yamabana, K., Muraki, K., Doi, S. & Kamei, S. (1996). A Language Conversion Front-end for Cross-Linguistic Information Retrieval. Working Notes of the Workshop on Cross-Linguistic Information Retrieval. Eds. G. Grefenstette & A. Smeaton & P. Sheridan. ACM SIGIR, Zürich, s. 34-39.

Viite

1. INQUERY-tiedonhakujärjestelmä on kehitetty USA:ssa Massachusettsin yliopiston Tietojenkäsittelytieteen laitoksen (Computer Science Department) Tiedonhaun laboratoriossa (Information Retrieval Laboratory)

Liite

Hakutehtävä 216 Osteoporoosi

Alla olevat 5 kyselyä (sd —> gd -käännös-

menetelmä) edustavat kaikkia tutkimuksen kyselytyyppejä. Rakenteeton ja rakenteinen LL CLIR-kysely ovat vertailukelpoisia. Huomaa kuinka hakuavaimella osteoporoosi, joka selvästikin on hakupyynnön "osteoporoosi" (ks. s. 56) tärkein ilmaisu, on suhteellisesti enemmän painoa rakenteissa LL CLIR-kyselyssä.

LL vertailu

#sum(what research is ongoing to reduce the effects of osteoporosis and prevent the disease)

Rakenteeton LL CLIR

#sum(osteoporosis prevention repression restraining restraint obstruction contraception dwarfing stunting disease #uw3(bring about) cause create effect #uw3(give rise to) inflict produce consequence effect outgrowth result lieventäminen join #uw3(join in) ally #uw3(join together) unite #uw3(be connected) #uw3(be linked) belong examination exploration inquest investigation report research scrutiny study meneillä)

Rakenteinen LL CLIR

#sum(osteoporosis #syn(prevention repression restraining restraint obstruction contraception dwarfing stunting) disease #syn(#uw3(bring about) cause create effect #uw3(give rise to) inflict produce) #syn(consequense effect outgrowth result) lieventäminen #syn(join #uw3(join in) ally #uw3(join together) unite #uw3(be connected #uw3(be linked) belong) #syn(examination exploration inquest investigation report research scrutiny study) meneillä)

KA vertailu

#and(osteoporosis #syn(prevent reduce) research)

KA CLIR

#and(osteoporosis #syn(prevent avert obviate obstruct hinder impede arrest delay retard avoid alleviate mitigate reduce weaken abate relieve ease lighten) #syn(examination exploration inquest investigation report research scrutiny study))