# Tilastollinen kääntäminen kieltenvälisessä tekstitiedonhaussa

Tuomas Talvensaari
Tampereen yliopisto
Tietojenkäsittelytieteiden laitos
Tuomas.Talvensaari@cs.uta.fi

## 1 Johdanto

Tiedonhaun (engl. *information retrieval, IR*) tutkimuksessa pyritään kehittämään menetelmiä relevantin tiedon löytämiseen erilaisista tietoalkioista, esimerkiksi tekstidokumenteista [8]. Perinteisessä tekstitiedonhakututkimuksessa on keskitytty tilanteeseen, jossa käyttäjän muotoilema kysely ja tietoalkiot sisältävä dokumenttikokoelma on ilmaistu samalla kielellä. Kieltenvälisessä tiedonhaussa (engl. *cross-language information retrieval, CLIR*) kysely ja haettavat dokumentit ovat erikielisiä. Kyselyn kieltä kutsutaan lähtökieleksi ja dokumenttien kieltä (tai kieliä) kohdekieleksi. Kieltenvälistä tiedonhakua tukevissa järjestelmissä käyttäjä voisi yhdellä lähtökielen kyselyllä hakea usean eri kielen dokumentteja [4].

Kyselyn ja dokumenttikokoelman välisen kielimuurin ylittämiseksi kysely yleensä käännetään kohdekielelle, minkä jälkeen suoritetaan normaali yksikielinen haku. Kyselyn kääntämiseen käytettävät menetelmät voidaan jakaa karkeasti kahtia sanakirjoihin perustuviin menetelmiin ja tilastolliseen kääntämiseen perustuviin menetelmiin. Sanakirjakääntämisessä käytetään hyväksi sähköiseen muotoon muunnettuja kieltenvälisiä sanakirjoja. Lähtökielen kyselysana yksinkertaisesti vaihdetaan käännösvastineisiinsa [6].

Tilastollinen kääntäminen puolestaan perustuu laajoihin monikielisiin dokumenttikokoelmiin. Käytettävät kokoelmat on usein rinnastettu dokumenttien tasolla siten, että kullekin lähtökielen dokumentille on kohdekielen dokumenttikokoelmassa pari. Rinnakkaiskokoelmissa (engl. *parallel collection*) dokumenttiparit ovat toistensa tarkkoja käännöksiä. Vastindokumenttikokoelmissa (engl. *comparable collection*) parit eivät ole käännöksiä, mutta dokumentit vastaavat toisiaan aiheiltaan, eli ne kertovat esimerkiksi samasta uutistapahtumasta. Koska rinnastetut dokumentit kertovat samoista asioista, voidaan olettaa, että dokumentit sisältävät yhtenevää sanastoa — esimerkiksi jos suomen- ja ruotsinkieliset dokumentit käsittelevät Venäjän tapahtumia, löytyy lähtödokumentista todennäköisesti useaan kertaan sana *Venäjä* ja kohdedokumentista sana *Ryssland* [7].

Tilastollisen kääntämisen tarkkuuden kannalta laaja rinnakkaiskokoelma olisi ihanteellinen. Tällaiset kokoelmat ovat kuitenkin melko harvinaisia, joten usein joudutaan luomaan vastindokumenttikokoelma kahdesta erillisestä tekstidokumenttikokoelmasta. Omassa tutkimukses-

sani olen pyrkinyt kehittämään menetelmän, jonka avulla kahdesta erillisestä, yksikielisestä kokoelmasta pystyttäisiin luomaan vastindokumenttikokoelma, sekä käyttämään menetelmän avulla luotuja kokoelmia tilastollisen kääntämisen apuvälineenä. Menetelmä esiteltiin lähteessä [9].

## 2 Dokumenttien pariutus

Olen tutkinut tilastollista kääntämistä useilla eri kielipareilla, joista eräs on ruotsi-englanti. Tässä kieliparissa lähtökokoelmana käytettiin ruotsinkielistä TT-uutistoimiston artikkeleista koostuvaa dokumenttikokoelmaa. Dokumentteja oli 142 819 kappaletta, ja ne oli toimitettu vuosina 1994 ja 1995. Kohdekokoelma käsitti 113 005 Los Angeles Times lehden vuonna 1994 ilmestynyttä artikkelia. Kokoelmat ovat osa kansainvälisen *Cross-Language Evaluation Forum* -konferenssin (*CLEF* [5]) testikokoelmaa.

Lähtökokoelman dokumentit pariutettiin kohdekokoelman dokumenttien kanssa. Kustakin lähtödokumentista eroteltiin parhaat hakusanat — siis sanat, jotka parhaiten kuvaavat dokumentin aihetta — analysoimalla sanojen frekvenssejä dokumenteissa. Nämä sanat käännettiin kohdekielelle UTACLIR-kyselynkäännöskoneella [3], joka perustuu sanakirjakääntämiseen. Näin saatiin kohdekielinen kysely, joka ajettiin kohdekokoelmaa vasten InQuery-hakukoneella [1]. Hakutuloksen kärkipäästä haettiin dokumenttia, joka olisi julkaistu muutaman päivän sisällä lähtödokumentin julkaisupäivämäärästä — näin pyrittiin varmistamaan se, että dokumenttipari kertoisi samasta tapahtumasta. Jos tällainen dokumentti löytyi, muodostettiin pari.

Kaikille lähtödokumenteille ei siis löydetty vastinparia, mikä onkin ymmärrettävää: läheskään kaikkia ruotsalaisen uutistoimiston raportoimia uutisia ei noteerata amerikkalaisessa sanomalehdessä. Edellä kuvatulla menetelmällä luotiin 9794 dokumenttiparista muodostuva vastindokumenttikokoelma.

## 3 Dokumenttiparien hyödyntäminen

Vastindokumenttikokoelman hyödyntämisessä käytettiin tiedonhaun ns. vektorimallia [2], jossa dokumenteissa esiintyvät sanat (tai itse dokumentit) voidaan esittää vektoreina. Vektoreiden alkiot ovat sanojen painoarvoja kokoelman dokumenteissa: mitä useammin sana *s* esiintyy dokumentissa *d*, sen suurempi on dokumentin *d* painoarvo sanaa *s* mallintavassa vektorissa. Kahden sanan keskinäistä samanlaisuutta voidaan mitata laskemalla vaikkapa vektoreiden kosinitulo. Vastindokumenttikokoelmassa voidaan mitata lähtökokoelman sanan ja kohdekokoelman sanan samanlaisuutta. Kun lähtökokoelman sanaa verrataan kaikkiin kohdekokoelman sanoihin, pitäisi samanlaisuusvertailun kärkeen sijoittua lähtösanaa semanttisesti lähellä olevia sanoja.

Luomaamme vastindokumenttikokoelmaa käytettiin edellä kuvatulla tavalla lähtökielen sanojen “kääntämiseen”. Samanlaisuuden mittaamiseen käytettiin erästä kosinitulon muunnelmaa. Taulukossa 1 on esitetty järjestelmän (*COCOT, Comparable Corpus Translation*) antamat samanlaisuuslaskelmat neljälle ruotsin kielen sanalle. Oikeat käännökset on tummennettu. Tilastollinen kääntäminen toimi parhaiten sanoilla, joilla on selvä merkitys, kuten yleis- ja erisnimillä (*rysk, barn*). Sen sijaan esimerkiksi verbit (*draga*) eivät kääntyneet yhtä hyvin.

| | barn | | rysk | | Tjetjenien | | draga | |
|---|---|---|---|---|---|---|---|---|
| 1 | **child** | 12.51 | **Russian** | 22.14 | **Chechnya** | 27.20 | support | 4.65 |
| 2 | find | 7.42 | Russia | 19.09 | Grozny | 25.25 | peace | 4.35 |
| 3 | family | 7.40 | Moscow | 17.47 | Dudayev | 25.07 | clear | 4.20 |
| 4 | life | 6.57 | Yeltsin | 15.18 | Chechen | 24.22 | talk | 4.09 |
| 5 | woman | 6.42 | soviet | 13.96 | Dzhokar | 19.18 | Clinton | 3.90 |
| 6 | live | 6.33 | Boris | 13.01 | Russian | 17.06 | control | 3.88 |
| 7 | year-old | 6.32 | russ | 11.54 | Chechens | 17.03 | war | 3.87 |
| 8 | found | 6.29 | military | 9.98 | Caucasus | 16.65 | area | 3.86 |
| 9 | mother | 6.25 | Kremlin | 9.72 | Kremlin | 15.73 | secretary | 3.84 |
| 10 | kill | 6.16 | republic | 9.22 | Moscow | 13.61 | organization | 3.77 |

Taulukko 1: COCOTin samanlaisuuslaskelmat neljälle ruotsin kielen sanalle

On huomionarvoista, että vaikka esimerkiksi *Grozny* ei ole sanan *Tjetjenien* oikea käännös, se on kuitenkin semanttisesti lähellä lähtösanaa. Näin se voitaisiin huoletta laittaa Tshetsheniaa koskevaan hakulauseeseen. Tällainen "semanttinen kyselynlaajennus" onkin yksi tilastollisen kääntämisen eduista verrattuna esimerkiksi sanakirjakääntämiseen.

COCOTia kokeiltiin yhdessä sanakirjoihin perustuvan UTACLIRin kanssa. Käännöskoneiden käyttöjärjestystä ja COCOTin parametreja varioitiin. Yhdistelmän toimivuutta todennettiin perinteisellä tiedonhaun ns. laboratoriomallin [2] menetelmillä, jotka perustuvat pitkälti saannin ja tarkkuuden laskemiseen. Saanti tarkoittaa hakutuloksessa olevien relevanttien dokumenttien osuutta kaikista relevanteista dokumenteista. Tarkkuus taas kertoo löydettyjen relevanttien dokumenttien osuuden kaikista löydetyistä dokumenteista. Ihannetapauksessa sekä saanti että tarkkuus ovat 100 %, jolloin on löydetty kaikki relevantit dokumentit eikä yhtään epärelevanttia dokumenttia.

Laboratoriomallissa luodaan joukko eri aihealueita koskevia hakutehtäviä, joihin haetaan relevantit dokumentit käytettävistä testikokoelmista. Hakutehtävistä muokataan jollain kokeiltavalla menetelmällä kyselyitä. Menetelmän tehokkuutta mittaavat tarkat saanti- ja tarkkuusluvut pystytään laskemaan, koska relevantit dokumentit tunnetaan.

Eräs tässä tutkimuksessa käytetyistä hakutehtävistä oli ruotsiksi "*Vilka skäl finns det för den ryska militära interventionen i Tjetjenien?*" ("Mitä syitä on Venäjän interventiolle Tshetsheniassa?"). Lähtökielen kyselysanoiksi otettiin sanat *intervention*, *skäl*, *rysk*, *Tjetjenien*, *militär* ja *finna*. Nämä käännettiin ensin UTACLIRillä, jonka käännökset näkyvät alla olevassa taulukossa.

| lähtösana | UTACLIR |
|---|---|
| intervention | intervention |
| skäl | dish, rind, peeling |
| militär | serviceman, military |
| finnas | find |

Sanat *rysk* ja *Tjetjenien* eivät olleet UTACLIRin käyttämässä sanakirjassa, joten ne käännettiin COCOTilla, joka palautti samanlaisuuslaskelman kolme parasta sanaa (ks. Taulukko 1). Käännetyillä sanoilla tehtiin haku englanninkieliseen testikokoelmaan käyttäen InQueryhakukonetta. Kaikki relevantit dokumentit mahtuivat hakutuloksen kärkipäähän, joten haku onnistui erittäin hyvin.

Kaiken kaikkiaan edellisen kaltaisia hakutehtäviä oli testeissä 91. Yksittäisten hakutehtävien saanti- ja tarkkuusarvoista laskettiin keskiarvot, jotka indikoivat, että yhdistetty UTACLIR-COCOT -kääntäminen on tehokkaampaa kuin UTACLIRin käyttäminen yksinään. COCOTin kehittämistä on jatkettu tässä kuvattujen testien jälkeen [10].

# Viitteet

[1] James Allan, James P. Callan, W. Bruce Croft, Lisa Ballesteros, Donald Byrd, Russell C. Swan ja Jinxi Xu: INQUERY at TREC-5. *The Fifth Text Retrieval Conference ,TREC-5*, ss. 119-132, Gaithersburg, Maryland, 1997. NIST Spec. pub. 500–238. National Institute of Standards and Technology, Gaithersburg, Maryland.

[2] Ricardo Baeza-Yates ja Berthier Ribeiro-Nieto: *Modern Information Retrieval*. Addison-Wesley, 1999.

[3] Heikki Keskustalo, Turid Hedlund ja Eija Airio: UTACLIR - general query translation framework for several language pairs. *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ss. 448–448, Tampere, 2002. ACM Press, New York.

[4] Douglas W. Oard ja Anne R. Diekema: Cross-language information retrieval. *Annual review of Information Science and Technology (ARIST)* 33:223–256, 1998.

[5] Carol Peters: What happened in CLEF 2004? Introduction to the working notes. http://www.clef-campaign.org/2004/working_notes/WorkingNotes2004/CLEF2004WN%20-%20intro.pdf (haettu 8.5.2006). Instituto di Scienza e Tecnologie dell'Informazione (ISTI-CNR), Pisa, Italia.

[6] Ari Pirkola, Turid Hedlund, Heikki Keskustalo ja Kalervo Järvelin: Dictionary-based cross-language information retrieval: problems, methods, and research findings. *Information Retrieval*, 4:209–230, 2001.

[7] Paraic Sheridan ja Jean-Paul Ballerini: Experiments in multilingual information retrieval using the SPIDER system. *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ss. 58–65, Zürich, 1996. ACM Press, New York.

[8] Amit Singhal: Modern information retrieval: a brief overview. *IEEE Data Engineering Bulletin*, 24(4):35–43, 2001.

[9] Tuomas Talvensaari, Jorma Laurikkala, Kalervo Järvelin ja Martti Juhola: A study on automatic creation of a comparable document collection in cross-language information retrieval. *Journal of Documentation*, 62(3):372–387, 2006.

[10] Tuomas Talvensaari, Jorma Laurikkala, Kalervo Järvelin ja Martti Juhola: Corpus-based CLIR in retrieval of highly relevant documents. *Journal of the American Society of Information Science and Technology (JASIST)* (to appear).