Informaatiotutkimuksen päivät 2010
21. - 22. lokakuuta, Tampere

ABSTRAKTI

Kimmo Kettunen

Tavut sananmuotojen vaihtelun hallinnan välineinä tekstitiedonhaussa

Kimmo Kettunen, Kymenlaakson ammattikorkeakoulu, kimmo.kettunen@kyamk.fi

Yksi tekstitiedonhaun ongelmista erityisesti kokotekstihaussa on kautta aikojen ollut hakutermeissä ja tekstidokumenteissa esiintyvä sananmuotojen vaihtelu. Eri kielet käyttäytyvät tässä suhteessa eri tavoin, esimerkiksi englannin kielessä sanat eivät muodoltaan juurikaan vaihtele, mutta suomessa sanojen muoto-oppi eli morfologia on rikasta. Tämä puolestaan vaikuttaa hakutuloksiin, jos sananmuotojen vaihtelulle ei tekstien haku- ja indeksointivaiheessa tehdä jotain. Perinteisiä keinoja sananmuotojen vaihtelun hallintaan ovat olleet hakutermien katkaisu, karsinta eli stemmaus sekä perusmuotoistaminen eli lemmaus. Kaikkia näitä menetelmiä on käytetty menestyksekkäästi erilaisissa tekstihakujärjestelmissä, ja erityisesti karsinta ja perusmuotoistaminen ovat muodostuneet hakujärjestelmien vakiomenetelmiksi. Monia muitakin menetelmiä on olemassa, ja erilaisia menetelmiä esittelevät kattavasti esimerkiksi Kettunen (2009) ja McNamee, Nicholas ja Mayfield (2009).

Yksi vähälle huomiolle jäänyt menetelmä sananmuotojen vaihtelun käsittelyyn on tavujen käyttäminen. Puheteknologiassa ja puhehaussa tavuja on käytetty paljonkin (ks. esimerkiksi Wang, 2000; Ng ja Zue, 2000), mutta kokotekstihaussa tavujen käyttäminen sananmuotojen vaihtelun hallinnan menetelmänä näyttää olleen vähäistä. Joitain poikkeuksiakin on, esimerkiksi Larson ja Eickeler (2003) esittelevät saksan kielen hakutuloksia, joissa tavutusta on käytetty menestyksekkäästi sekä puhutun että kirjoitetun dokumenttiaineiston haussa. Tavunkaltaisia, partikkeleiksi kutsumiaan elementtejä, käyttävät puolestaan Gouvea ja Raj (2009). Kirjoittajat tosin myös toteavat, että heidän partikkelinsa eivät ole tavuja, vaikka muistuttavatkin niitä.

Tavut

Tavuja pidetään kielitieteessä yhtenä kielen perusrakenteena. Kielen rakenteessa tavut kuuluvat fonologiaan eli äänneoppiin, jos kohta voidaan puhua myös ortografisen eli oikeinkirjoitustason tavutuksesta. Kussakin kielessä on melko rajallinen määrä sallittuja abstrakteja tavutyyppejä, jotka toteutuvat rajallisena määränä konkreettisia tavuja. Esimerkiksi Karlsson (1982) esittää, että suomessa on kymmenen kielen ytimeen kuuluvaa tavutyyppiä ja noin 3000 erilaista konkreettista tavua. Muiden kielten konkreettisia tavumääriä esitetään esimerkiksi lähteessä Pellegrino, Coupé ja Marsico (2007). Vähiten konkreettisia tavuja Pellegrinon ja kumppaneiden luettelossa on kiinassa ja japanissa, 416, eniten englannissa, 7931.

Tavutus voi vaikuttaa yksinkertaiselta prosessilta, mutta itse asiassa se ei ole sitä, koska tavun kielitieteellisessä määrittelyssä on ongelmia. Tästä syystä myös eri kielissä voi esiintyä

erilaisia tavutuksia samoille sanoille. Samasta syystä myös automaattisten tavuttimien tarkkuus vaihtelee. Parhaat automaattiset tavuttimet pääsevät yli 99 prosentin kattavuuteen (ks. esimerkiksi Bouma, 2003), mutta monien tavuttimien kattavuus jää 95-96 prosenttiin. Erilaisia tavutusalgoritmeja esittelevät muun muassa Bouma (2003), Adsett ja Marchand (2009), Marchand, Adsett ja Damper (2007) sekä Bartlett, Kondrak ja Cherry (2007). Esitellyt tavuttimet ovat osaksi sääntöpohjaisia, toisin sanoen ihminen on kirjoittanut tavuttimen säännöt, osittain ns. oppivia ohjelmia, jotka päättelevät annetusta malliaineistosta tarvittavat säännöt itse. Tavutuksen laatu vaihtelee kielestä ja lähestymistavasta toiseen, mutta yleisesti erilaiset oppivat menetelmät näyttävät selviävän tavutuksessa hyvin.

Aineistot ja hakuympäristöt

Päätimme tutkia tavutuksen käyttöä tekstitiedonhaussa sananmuotojen vaihtelun hallinnassa. Lähtökohtana tutkimukselle oli tieto siitä, että n-grammit, eli tekstien sanoista muodostetut 2-6 merkkijonon mittaiset pätkät ovat toimineet erittäin hyvin tekstihaussa (esim. McNamee, 2008; McNamee ja Mayfield, 2004). N-grammien keskeinen ongelma hakukäytössä on kuitenkin se, että niitä käytettäessä hakuindekseistä muodostuu erittäin suuria ja haut ovat hitaita. Jos sanat ositetaan tavuihin, syntyy osin samanlaisia merkkijonoja kuin n-grammauksessa, mutta merkkijonojen määrä jää huomattavasti pienemmäksi.

Käytimme tutkimuksessa 14 kielen tekstidokumenttikokoelmia. 13 kielistä on Cross Language Forumin (CLEF) aineistoja, neljästoista turkkia, josta on julkaistu tuore Milliyet-sanomalehtitekstikanta (Can ja kumppanit, 2008). CLEF-aineistojen tiedonhakutestit teki Paul McNamee Johns Hopkins –yliopistosta käyttäen yliopistossa kehitettyä HAIRCUT-hakujärjestelmää (McNamee ja Mayfield, 2004). Turkin hakukokeet teki Tampereen yliopiston Informaatiotutkimuksen ja interaktiivisen median laitoksella Feza Baskaya Lemur-hakujärjestelmällä (Lemur). Tarkemmat aineistotiedot on esitetty artikkelissa Kettunen, McNamee ja Baskaya (2010).

Tulokset

Päätimme ottaa lähtökohdaksi hakutesteihin ensin hyvin yksinkertaisen lähestymistavan. Lähes jokaisessa kielessä esiintyy yhtenä perustavutyyppinä CV, eli konsonantista ja vokaalista muodostuva tavu. Onpa kieliä, joissa tällainen tavu on ainoa sallittu tavutyyppi (Maddieson, 2008), ja erään teorian mukaan kaikkien kielten kaikki tavut ovat itse asiassa CV-tavuja (van der Hulst ja Ritter, 1999). Niinpä sovelsimme ensin kuhunkin kieleen kahta erilaista tavutusta, jossa tavunraja sijoitettiin joko jokaisen CV-jakson jälkeen tai ennen sitä. Jälkimmäinen tavutustapa toimii esimerkiksi suomessa perustavana tavusääntönä, jolla on suuri kattavuus (Karlsson, 1985). Kielestä riippuen nämä prosessit tuottavat tuloksena luonnollisesti sekä oikeita että vääriä tavuja, mutta tällä ei ole varsinaisesti merkitystä lopputuloksen kannalta. Voisimmekin itse asiassa puhua tavujen sijasta tavugrammeista, tavujen kaltaisista merkkijonoyhdistelmistä, joita käytetään hakutermien ja dokumenttien käsittelyssä. Näitä kahta yksinkertaista tavutustapaa kutsutaan jatkossa nimillä CV_1 ja CV_2. Ne tuottaisivat seuraavat tavutukset esimerkkisanoista: CV_1 (*ca + rbo + hy + dra + te + s; do + gs; go + es) CV_2 (car+bo+hyd+ra+tes; dogs; goes).*

Kun dokumenttiaineistosta tehtiin hakuindeksit, luotiin tavutusta käyttäen kolme erilaista indeksiä: 1) yksittäisten tavujen indeksit, 2) kahden tavun yhdistelmien indeksit ja 3) kolmen tavun yhdistelmien indeksit. Vastaavasti kyselyitä suoritettaessa kyselyistä muodostettiin samanlaiset merkkijonoyhdistelmät. Taulukossa 1 esitetään tulokset kaikille 14 kielelle CV_1 ja CV_2-tavutusta käyttäen hakujen keskitarkkuuksina.

Taulukko 1. Keskitarkkuudet CV_1 ja CV_2-tavutuksella. Otsikko- ja kuvailukenttäkyselyt

	words	**snow**	**4**	**syl1_ CV1**	**syl2_ CV1**	**syl3_ CV1**	**syl1_ CV2**	**syl2_ CV2**	**syl3_ CV2**
BG	0.22	N/A	**0.31**	0.21	0.22	0.10	0.21	0.20	0.10
CS	0.23	N/A	**0.33**	0.18	0.26	0.16	0.19	0.27	0.19
DE	0.33	0.37	**0.41**	0.28	0.39	0.29	0.30	0.38	0.24
EN	0.41	0.44	**0.40**	0.21	0.38	0.27	0.23	0.35	0.20
ES	0.44	0.49	**0.46**	0.24	0.45	0.31	0.22	0.43	0.29
FI	0.34	0.43	**0.50**	0.30	0.46	0.38	0.27	0.43	0.31
FR	0.36	**0.40**	0.38	0.20	0.37	0.25	0.23	0.34	0.22
HU	0.20	N/A	**0.38**	0.20	0.32	0.23	0.18	0.29	0.18
IT	0.38	**0.42**	0.37	0.18	0.39	0.26	0.17	0.37	0.26
NL	0.38	0.40	**0.42**	0.26	0.38	0.25	0.29	0.36	0.23
PT	0.32	N/A	**0.34**	0.17	0.33	0.20	0.17	0.30	0.16
RU	0.27	N/A	**0.34**	0.28	0.24	0.13	0.26	0.26	0.15
SV	0.34	0.38	**0.42**	0.26	0.41	0.31	0.25	0.37	0.26
TU	0.19	0.22	**0.31**	0.17	0.30	0.22	0.21	0.26	0.20

Taulukon selite: words = juoksevat sananmuodot sellaisinaan (tämä on testien perustaso); snow = Snowball-karsimella karsitut sananmuodot; 4 = neligrammit ; syl1, syl2, syl3 = yhden, kahden ja kolmen tavun yhdistelmät

Taulukossa 2 on esitetty muutokset perustasoon.

Taulukko 2. Muutokset

	words	snow	4	syl1_ CV1	syl2_ CV1	syl3_ CV1	syl1_ CV2	syl2_ CV2	syl3_ CV2
Avg-8	0.37	0.42	0.42	0.24	0.40	0.29	0.25	0.38	0.25
Chg-8 %	N/A	11.47	13.31	-34.69	7.89	-22.18	-34.14	1.60	-32.80
Avg-A	0.32	N/A	0.39	0.23	0.35	0.24	0.23	0.33	0.21
Chg-A %	N/A	N/A	20.54	-29.15	8.69	-25.25	-29.42	3.40	-33.75

Taulukon selite: Avg-8 = kahdeksan Snowballilla käsitellyn kielen keskiarvo; Chg-8 % = Snowball-kielten tulosten keskiarvon suhteellinen prosentuaalinen muutos perustasoon nähden; Avg-A = kaikkien CLEF-kielten keskiarvo; Chg-A % = kaikkien CLEF-kielten keskiarvon suhteellinen prosentuaalinen muutos perustasoon nähden.

Tuloksista ilmenee, että 4-grammit tuottivat parhaita keskitarkkuuksia lähes jokaisessa kielessä. Kahden tavun yhdistäminen indeksoinnissa ja hakutermien käsittelyssä tuotti myös hyviä tuloksia monissa kielissä, kun käytettiin CV_1-tavutusta. Kiintoisinta on havaita, että useissa kielivalikoiman morfologisesti mutkikkaissa kielissä (suomi, ruotsi, saksa, turkki ja unkari) tavutus toimi varsin hyvin. Keskitarkkuudet eivät jääneet paljon 4-grammien tarkkuuksista, ja myös Snowball-karsinnan tarkkuuden tasolle päästiin tai se jopa hiukan ylitettiin. Snowballilla tehdyn sanojen karsinnan (stemming) on erilaisissa kokeissa todettu tuottavan hyviä tuloksia (Airio, 2006), joten samalle tarkkuustasolle sen kanssa pääseminen erityisesti morfologisesti mutkikkaammissa kielissä on jo hyvä suoritus. CV_2-tavutus ei toiminut aivan yhtä hyvin kuin CV_1, mutta myös sillä saavutettiin kohtalaisia tuloksia.

N-grammien hyvä ja tasainen suoritus lähes kaikissa kielissä ei yllätä, tämä on osoitettu moneen kertaan aiemminkin. Niiden hyvän suorituksen hinta on kuitenkin melko kova: indeksien koko voi olla moninkertainen verrattuna käsittelemättömiin sanoihin ja haut ovat myös hitaita. McNameen, Nicholasin ja Mayfieldin (2009) kokeissa englannin kielen indeksien koko kasvoi kolminkertaiseksi ja hakujen tekeminen oli seitsemän kerta hitaampaa kun verrattiin käsittelemättömiä sanoja ja 4-grammeja.

Koska käyttämämme tavutukset, CV_1 ja CV_2, olivat molemmat melko suoraviivaisia ja yksinkertaistavia, halusimme testata myös millaisia tuloksia syntyy, jos tavuttimen laatu on parempi. Meillä oli käytössä kunnon tavutusohjelmat kolmelle kielelle: saksalle, suomelle ja turkille[1]. Suomessa yhden, kahden ja kolmen tavun yhdistelmillä saavutettiin keskitarkkuudet 0.28, 0.44 ja 0.33. Saksassa vastaavat tulokset olivat 0.31, 0.36 ja 0.23 ja turkissa 0.21, 0.27, ja 0.20. Tulokset jäivät siis jonkin verran CV_1-prosessin parhaista tuloksista, mutta kahden tavun yhdistelmän tuloksia ei tässäkään tapauksessa voida pitää huonoina. Suomessa ja turkissa kunnollisen tavutuksen tulokset ylittivät myös Snowballin tulokset, kun käytettiin kahden tavun yhdistelmiä.

Kovapintainen tiedonhakuteoreetikko saattaa tässä vaiheessa kohauttaa olkapäitään ja todeta, että tekeekö tällä sananmuotojen vaihtelun hallinnan lähestysmistavalla yhtään mitään, eihän se kykene parempaan suoritukseen kuin 4-grammit. Kuten on aiemmin todettu, 4-grammit ovat erittäin tehokkaita, mutta ne vievät myös erittäin paljon resursseja: indeksit levytilaa ja kyselyt hakuaikaa. Niitä tuskin kukaan käyttää tuotantokäytössä olevissa hakujärjestelmissä näistä syistä. Yksinkertainen CV-tavutus puolestaan on helppo toteuttaa, eivätkä indeksit muodostu tavattoman suuriksi. Hakunopeutta emme ole testanneet toistaiseksi, mutta se lienee ainakin kohtuullinen. Kiintoisaa on myös se, että karkeapiirteinen CV-tavutus toimii monessa kielessä hyvin. Menetelmää ei voi sanoa kieliriippumattomaksi, mutta vähintäänkin se on kielen suhteen joustava. On todennäköistä, että erilaiset kielitypologiset syyt vaikuttavat siihen, missä kielessä CV-tavutus toimii hyvin ja missä ei (Fenk-Oczlon ja Fenk, 1999). Tätä puolta olisi myös mielenkiintoista tutkia, mutta se veisi teemaa liian kauaksi tiedonhausta.

Kirjallisuutta

Addsett, C., Marchand, Y. (2009). A Comparison of data-driven automatic syllabification methods. String Processing and Information Retrieval, 16th International Symposium, SPIRE 2009 (toim. Jussi Karlgren, Jorma Tarhio & Heikki Hyyrö), s. 174–181. Heidelberg: Springer.

Airio, E. (2006). Word normalization and decompounding in mono- and cross-lingual IR. Information Retrieval 9: 249–271.

Bartlett, S., Kondrak, G., Cherry, C. (2008). Automatic syllabification with structured SVMs for letter-to-phoneme conversion. Proceedings of ACL-08, HLT, s. 568–576. Columbus.

[1] Suomen kielen tavutin on artikkelin kirjoittajan toteuttama, saksan ja turkin tavuttimet Feza Baskayan tekemiä.

Bouma, G. (2003). Finite state methods for hyphenation**.** Natural Language Engineering 9: 5–20

Can, F., Kocberber, S., Balcik, E., Kaynak, C., Ocalan, H. C., Vursavas, O. M. (2008). Information retrieval on Turkish texts. Journal of the American Society for Information Science and Technology 59: 407–421.

Fenk-Oczlon, G., Fenk, A. (1999). Cognition, quantitative linguistics, and systemic typology. Linguistic Typology 3:151–177.

Gouvea, E. B., Raj, B. (2009). Word particles applied to information retrieval. Advances in information retrieval. 31th European Conference on IR Research, ECIR 2009. (toim. Mohand Boughanem, Catherine Berrut, Josiane Mothe, Chantal Soule-Dupuy), s. 424–436. Heidelberg: Springer.

Van der Hulst, H., Ritter, N. A. (1999). The syllable: views and facts. Berlin: Mouton de Gruyter.

Karlsson, F. (1982). Suomen kielen äänne- ja muotorakenne. Porvoo: WSOY.

Karlsson, F. (1985). Automatic Hyphenation of Finnish. (toim. Fred Karlsson) Computational morphosyntax. Report on Research 1981–1984, s. 93–113 Publications of the Department of General Linguistics, University of Helsinki, 13.

Kettunen, K. (2009). Reductive and generative approaches to management of morphological variation of keywords in monolingual information retrieval – an overview. Journal of Documentation 2: 267–290.

Kettunen, K., McNamee, P., Baskaya, F. (2010). Using Syllables as Indexing Terms in Full-text Information Retrieval. HLT 2010, Riga. Ilmestyy.

Larson, M., Eickeler, S. (2003). Using syllable-based indexing features and language models to improve German spoken document retrieval. Proceedings of Eurospeech. 8th European Conference on Speech Communication and Technology. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.124.4455&rep=rep1&type=pdf. 3.9.2010.

Lemur. http://www.lemurproject.org/ 3.9.2010.

Maddieson, I. (2008). Chapter 12: Syllable Structure. The World Atlas of Language Structures Online. http://wals.info/feature/12. 3.9.2010.

Marchand, Y., Adsett, C., Damper, R. (2007). Evaluating automatic syllabification algorithms for English. http://eprints.ecs.soton.ac.uk/14285/1/MarchandAdsettDamper_ISCA07.pdf. 3.9.2010.

Marchand, Y., Adsett, C., Damper, R. (2009). Automatic syllabification in English: a comparison of different algorithms. Language and Speech 52: 1–27.

McNamee, P. (2008) Textual representations for corpus-based bilingual retrieval. PhD Thesis, University of Maryland Baltimore County. http://apl.jhu.edu/~paulmac/publications/thesis.pdf. 3.9.2010.

McNamee, P., Mayfield, J. (2004). Character n-gram tokenization for European language text retrieval. Information Retrieval 7: 73–97.

McNamee, P., Nicholas, C., Mayfield, J. (2009). Addressing morphological variation in alphabetic languages. Proceedings of the 32nd Annual International Conference on Research and Development in Information Retrieval (SIGIR-2009), s. 75–82. Boston, MA.

Ng, K., Zue, V. W. (2000). Subword-based approaches for spoken document retrieval. Speech Communication 32: 157–186.

Pellegrino, F., Coupé, C., Marsico, E. (2007). An Information theory-based approach to the balance of complexity between phonetics, phonology and morphosyntax. http://www.ddl.ish-lyon.cnrs.fr/fulltext/pellegrino/Pellegrino_2007_PCM_LSA.pdf. 3.9.2010.

Wang, H.-M. (2000). Experiments in syllable-based retrieval of broadcast news speech in Mandarin Chinese. Speech Communication 32: 49–60.