TEKNILLINEN KORKEAKOULU
Elektroniikan, tietoliikenteen ja automaation tiedekunta

Tommi Vatanen

KIELENTUNNISTUS LYHYISTÄ TEKSTILOHKOISTA N-GRAMMI-MALLEIHIN PERUSTUVALLA LUOKITTIMELLA

Kandidaatintyö

Espoo 7. joulukuuta 2009

Työn ohjaaja:

DI Jaakko Väyrynen

TEKNILLINEN KORKEAKOULU

KANDIDAATINTYÖN TIIVISTELMÄ

Tekijä: Tommi Vatanen

Työn nimi: Kielentunnistus lyhyistä tekstilohkoista n-grammimalleihin perustuvalla luokittimella

Päivämäärä: 7. joulukuuta 2009 Kieli: suomi Sivumäärä: 7+27

Tutkinto-ohjelma: Bioinformaatioteknologia

Vastuuopettaja: TkT Markus Turunen

Ohjaaja: DI Jaakko Väyrynen

Tässä opinnäytetyössä tutkittiin kokeellisesti kielentunnistusmenetelmien toimintaa lyhyillä, 5–21 merkin mittaisilla tekstilohkoilla. Työssä ehdotettu KNLM-menetelmä perustuu suurimman uskottavuuden luokittimeen, joka mallintaa kieliä Kneyser–Ney-tasoitetuilla ja karsituilla merkkipohjaisilla kielimalleilla. Kokeissa käytettiin aineistona YK:n ihmisoikeusjulistusta, josta oli käytössä käännökset 298 kielelle. Kokeellisessa osuudessa vertailtiin KNLM-menetelmän suoriutumista yksinkertaisempia kielimalleja käyttävään LLM-menetelmään (Dunning, 1994), yleisesti käytettyyn Ranking-menetelmään (Cavnar ja Trenkle, 1994) sekä Google AJAX Language APIin. KNLM-menetelmä osoittautui käyttökelpoiseksi lyhyiden tekstilohkojen kielen tunnistamiseen, ja se suoriutui testeistä kaikkia vertailumenetelmiä paremmin.

Avainsanat: kielentunnistus, n-grammimalli, tilastollinen kielimallinnus

Esipuhe

Haluan kiittää ohjaajaani Jaakko Väyrystä erittäin hyvästä ja asiantuntevasta ohjauksesta. Häneltä on ollut ilo saada ideoille ja ajatuksille sekä tukea että tarvittaessa myös kritiikkiä. Kiitän myös Sami Virpiojaa, joka on useaan otteeseen antanut minulle asiantuntevia neuvoja ja kommentoinut työni eri yksityiskohtia. Lisäksi kiitän kaikkia, jotka ovat lukeneet ja kommentoineet työtäni sen eri vaiheissa. Kiitos kuuluu myös Timo Honkelalle ja kaikille tutkimusryhmämme jäsenille kannustavasta ja motivoivasta työilmapiiristä, joka ryhmässämme vallitsee.

Espoo, 7. joulukuuta 2009

Tommi Vatanen

Sisältö

Sanasto, lyhenteet ja symbolit

Sanasto

absoluuttinen alennus	absolute discounting
alivuoto	underflow
entropia	entropy
haarautumistekijä	branching factor
hämmentyneisyys	perplexity
interpolointi	interpolating
karsiminen	pruning
korpus, kieliaineisto	corpus, mon. corpora
k-kertaisesti ristiinvalidoitu t-testi	k-fold cross-validated paired t-test
merkkijonoydin	string kernel
n-grammi, n-piirre	n-gram
n-grammimalli	n-gram language model
n-grammioletus	n-gram assumption
nollahypoteesi	null hypothesis
ohjattu oppiminen	supervised learning
opetusaineisto	training data
perääntyminen	back-off
ristientropia	cross-entropy
ristiinvalidointi	cross-validation
sisäinen arviointi	intrinsic evaluation
tasoitus	smoothing
tukivektorikone	support vector machine (SVM)
validointiaineisto	held-out data
vaihtoehtoinen hypoteesi	alternative hypothesis
ylioppiminen	overfitting, overtraining

Lyhenteet

WWW	World Wide Web, Internet-verkossa toimiva hajautettu hypertekstijärjestelmä
MLE	Maximum likelihood estimate, suurimman uskottavuuden estimaatti
KN	Kneser–Ney (-tasoitus)
KP	Kneser pruning, Kneser-karsiminen
ML	Maximum likelihood, suurin uskottavuus
PPM	Prediction by Partial Match (tekstintiivistysmenetelmä)
LLM	Laplace Language Model, Laplace-tasoitettu kielimalli
KNLM	Kneser-Ney Language Model, Kneser–Ney-tasoitettu kielimalli

Symbolit ja operaattorit

c_i	tekstin i:s merkki
c_j^i	merkkijono $c_j \dots c_i$
$P(X)$	X:n todennäköisyys
$P(X\|Y)$	ehdollinen todennäköisyys X, siten että Y
$C(\cdot)$	merkkijonon esiintymiskertojen määrä
$P_{\text{Lap}}(\cdot)$	Laplace-tasoitettu kielimalli
$\|A\|$	joukon A koko
N_c	opetusaineistossa c kertaa tavattujen n-grammien määrä
c^*	Good–Turing-tasoituksen arvio luokkien lukumäärille
$P_{\text{bo}}(\cdot)$	perääntymistä käyttävä kielimalli
$P_{\text{interp}}(\cdot)$	interpolointia käyttävä kielimalli
$P_{\text{abs}}(\cdot)$	absoluuttista alennusta käyttävä kielimalli
D	alennusta käyttävien tasoitusmenetelmien alennustermi
$N(\cdot)$	erilaisten merkkijonojen määrä
$N(\bullet c_{i-n+2}^i)$	erilaisten kontekstien määrä: $\|\{c_{i-n+1} : C(c_{i-n+1}^i) > 0\}\|$
$P_{\text{KN}}(\cdot)$	Kneser–Ney-tasoitettu kielimalli
$PP(\cdot)$	hämmentyneisyys
O	opetusaineisto
C	mielivaltainen merkkijono
K	kieli
$P_K(\cdot)$	kielimalli kielelle K
K_{ML}	ML-luokittimella valittu kieli
$H(X)$	X:n entropia
$H_{\text{C}}(X, Y)$	X:n ja Y:n jakaumien välinen ristientropia
$\mathcal{X}$	kielimateriaali, korpus
$\mathcal{X}_i$	kielimateriaalin osa i
$X \setminus Y$	joukkojen X ja Y erotus (set theoretic difference)
H_0	nollahypoteesi
H_1	vaihtoehtoinen hypoteesi
t_n	Studentin t-jakauma vapausastein n
$t \sim t_n$	t noudattaa jakaumaa t_n

1 Johdanto

Luonnollisten kielten käsittelyssä esiintyy usein tarve tunnistaa käsiteltävänä olevan tekstin kieli. Kielen tunnistamista voidaan tarvita esimerkiksi konekäännöksessä, näppäimistön tai WWW-selaimen merkistön valinnassa, optisessa tekstinluvussa, oikoluvussa tai tiedonhaussa. Myös jatkuvasti kasvavien tietoaineistojen automaattisessa käsittelyssä ja tiedonlouhinnassa tarvitaan usein kielentunnistusta.

Kielentunnistusta on tutkittu jo pitkään sekä kirjoitetusta tekstistä (Gold, 1967) että puheesta (House ja Neuburg, 1977). Nykyään yleisesti käytössä olevien tekstipohjaisten kielentunnistimien toiminta voidaan jakaa kahteen vaiheeseen: (i) opetusvaiheeseen, jossa rakennetaan mallit jokaisesta tunnetusta kielestä opetusaineiston perusteella, ja (ii) tunnistusvaiheeseen, jossa vertaillaan tunnistettavaa tekstiä malleihin ja valitaan parhaiten tekstille sopiva malli. Koska näin suoritetussa kielentunnistuksessa tarvitaan opetusvaiheessa näyte kustakin kielestä, kielentunnistus on esimerkki *ohjatusta oppimisesta*. Mahdollisten kielten joukko täytyy määritellä etukäteen, joten kielentunnistus on myös *luokitteluongelma*.

Vaikka kielentunnistukseen on kehitetty lukuisia menetelmiä, on kirjallisuudesta erotettavissa kaksi paljon käytettyä lähestymistapaa: (i) tilastollisten kielimallien käyttö ja (ii) kielille yleisten sanojen tai muiden piirteiden frekvenssien vertailu (sanastopohjaiset menetelmät). Tunnistuksessa mukana olevista kielistä tarvitaan kummassakin lähestymistavassa opetuskorpus eli malliaineisto, joka toimii esimerkkinä kustakin kielestä.

Yksi ensimmäisistä merkkipohjaisia tilastollisia kielimalleja kielentunnistuksessa hyödyntäneistä tutkijoista oli Dunning (1994). Sanastoon perustuvat menetelmät (esim. Souter ym., 1994) puolestaan pyrkivät tunnistamaan kieliä niissä yleisesti esiintyvien sanojen perusteella. Näistä kahdesta lähestymistavasta merkkipohjaiset kielimallit suoriutuvat paremmin lyhyillä syötteillä, koska ne mallintavat kieliä sanoja yksityiskohtaisemmilla piirteillä. Řehůřek ja Kolkus (2009) ovat kehittäneet sanastoon pohjautuvan menetelmän, jonka he raportoivat toimivan kielimalleihin pohjautuvien menetelmien tasoisesti.

Aikaisempi kielen tunnistamisen tutkimus on keskittynyt pääasiassa kokonaisten dokumenttien tai lauseiden kielen tunnistamiseen. Esimerkiksi Teahan (2000, 2001) on testannut menetelmäänsä 10 000 merkin ja noin kilotavun mittaisilla teksteillä. Cavnar ja Trenkle (1994) käyttivät noin 1,7 kilotavun tekstejä. Eräiden tutkijoiden mielestä kokonaisten dokumenttien kielen tunnistaminen on ratkaistu ongelma (McNamee, 2005). Pitkänkin dokumentin kielen tunnistaminen saattaa kuitenkin olla ongelmallista, jos dokumentti on monikielinen tai halutaan tunnistaa esimerkiksi käytetty murre (da Silva ja Lopes, 2006).

Tässä tutkimuksessa on keskitytty lyhyiden, 5–21 merkin mittaisten merkkijonojen kielen tunnistamiseen. Merkkijonoja ei rajattu sisältämään pelkästään kokonaisia sanoja, vaan ne saattavat alkaa tai loppua mielivaltaisesti keskeltä sanoja tai koostua pelkästään osasta sanaa.[1] Näistä lyhyistä merkkijonoista käytetään jatkossa nimeä tekstilohko. Kokonaisia dokumentteja hyvin tunnistavat menetelmät, kuten sanastopohjaiset menetelmät, eivät välttämättä suoraan sovellu tekstilohkojen kielen tunnistamiseen. Lisäksi lyhyitä tekstilohkoja tunnistettaessa voidaan olettaa, että koko tekstilohko on yksikielinen.

Toinen tässä tutkimuksessa korostettu tekijä on ollut mahdollisimman suuri kielten määrä. Sanastopohjaisten menetelmien vaatimuksena on, että kieli koostuu helposti toisistaan erotettavista sanoista. Näin ei kuitenkaan ole useiden ei-länsimaisten kielten kohdalla. Kielten määrä erottaa tämän työn aikaisemmasta tutkimuksesta, sillä kielentunnistusmenetelmien testaamiseen käytettyjen kielten määrä on ollut yleisesti suhteellisen pieni – usein alle kymmenen. Koska esimerkiksi EU:lla on tällä hetkellä 23 virallista kieltä, suppeat kielivalikoimat jättävät huomiotta monia reaalimaailman ongelmia.

Tutkimuksessa esitelty menetelmä käyttää kielten mallintamiseen puheentunnistuksessä käytetyillä menetelmillä laskettuja n-grammimalleja (Chen ja Goodman, 1999; Siivola ym., 2007). N-grammimalleille kehitetyt tasoitus- ja karsimismenetelmät ovat olleet tärkeitä puheentunnistuksen kehityksessä. Esimerkiksi Zissman (1996) on soveltanut n-grammimalleja myös kielen tunnistamiseen puheesta.

Tutkimuksen tarkoituksena on ollut vertailla neljän eri menetelmän toimintaa lyhyillä, 5–20 merkin mittaisilla tekstilohkoilla. Menetelmiä olivat työssä esitelty KNLM-menetelmä, Dunningin (1994) tilastollinen menetelmä, Cavnarin ja Trenklen (1994) Ranking-menetelmä ja Google AJAX Language API. Tutkimusaineistona käytettiin YK:n ihmisoikeusjulistusta, joka on saatavana yli 360 kielellä. Kokeet suoritettiin käyttäen kahta eri kielijoukkoa, jotka sisälsivät 60 ja 298 kieltä. Pienempi kielijoukko oli käytössä, jotta voitiin suorittaa vertailu myös suppeamman kielivalikoiman omaavaan Google AJAX Language APIn kanssa.

[1] Huomattakoon, että kaikki tutkimuksessa mukana olevat kielet eivät koostu länsimaalaisten kielten tapaan sanoista tai sisällä välilyöntejä erottamassa sanoja.

2 Menetelmät

Tässä luvussa esitellään työssä käytettyjä menetelmiä. Aluksi perehdytään *n-grammimalliin* (Chen ja Goodman, 1999; Goodman, 2008), joka on yksi tärkeimmistä luonnollisen kielen analysointiin käytettävistä työkaluista. Sen jälkeen esitellään n-grammimalleille käytettäviä *tasoitusmenetelmiä* (Chen ja Goodman, 1999) mukaan lukien nykyisin eniten käytetty Kneser–Ney-tasoitus.

Seuraavaksi tarkastellaan kieltä stokastisena prosessina informaatioteoreettista taustaa vasten. Taustaksi esitellään *entropian* (Shannon, 1948) käsite tilastollisen kielenmallinnuksen näkökulmasta. Lisäksi liitetään kielimallien evaluoinnissa käytetyt mitat *hämmentyneisyys* ja *ristientropia* (Rosenfeld, 1996) informaatioteoreettiseen viitekehykseen. Lopuksi esitellään kielentunnistimien vertailuun käytetyt menetelmät *ristiinvalidointi* (Alpaydin, 2004, ss. 330–331), *parittainen t-testi* (Alpaydin, 2004, ss. 342–343) ja *Fisherin menetelmä* p-arvojen yhdistämiseksi (Fisher, 1948) sekä kokeissa käytetyt kielentunnistusmenetelmät.

2.1 N-grammimalli

N-grammimallit ovat eräitä tärkeimmistä luonnollisen kielen käsittelyssä käytettävistä työkaluista. N-grammilla tarkoitetaan n merkin, morfeemin tai sanan mittaista jonoa. Yleisimmin käytettyjä malleja ovat sana- (Chen ja Goodman, 1999; Goodman, 2008) tai morfeemipohjaiset (Creutz ja Lagus, 2007) mallit, mutta tässä työssä on käytetty ainoastaan merkkipohjaisia n-grammimalleja. N-grammimallit ovat yksi tapa rakentaa *tilastollisia kielimalleja*.

Merkkipohjaisessa n-grammimallissa oletetaan, että kieli rakentuu johdonmukaisesta merkistöstä, johon kuuluvat kaikki kirjoitetussa kielessä esiintyvät merkit, mukaan lukien välimerkit ja -lyönnit sekä isot kirjaimet. Mallia luotaessa oletetaan, että jokaisen merkin esiintyminen riippuu ainoastaan $(n-1)$ edellisestä merkistä: (Goodman, 2008)

$$P(c_i|c_1^{i-1}) \approx P(c_i|c_{i-(n-1)}^{i-1}), \tag{1}$$

jossa c_1^{i-1} on tekstilohko $c_1 \dots c_{i-1}$. Tätä kutsutaan *n-grammioletukseksi* ja siinä oletetaan kielen käyttäytyvän kuten $(n-1)$ asteinen Markov-prosessi. Esimerkiksi 2-grammimalli (engl. *bigram model*) pyrkii ennustamaan uuden merkin sitä edeltäneen merkin perusteella:

$$P(\texttt{i}|\texttt{suomenkiel}) \approx P(\texttt{i}|\texttt{l}). \tag{2}$$

1-grammimalli (engl. *unigram model*) puolestaan vastaa yksittäisten merkkien todennäköisyysjakaumaa.

Merkkipohjainen kielimalli siis pyrkii ennustamaan tekstissä esiintyviä merkkejä $(n-1)$ edellisen merkin perusteella. Ennustamiseen läheisesti liittyvä tehtävä on mielivaltaisten tekstilohkojen todennäköisyyksien laskeminen kielimallin perusteella. Kielimallit perustuvat niiden luomisessa käytettyihin *korpuksiin* eli kieliaineis-

toihin. Merkkipohjaisten mallin tapauksessa *opetusaineistona* käytettävästä korpuksesta lasketaan merkkejä ja n-grammeja aina haluttuun pituuteen saakka.

Kieliaineistosta laskettujen frekvenssien perusteella voidaan arvioida eri merkkien todennäköisyyksiä (Goodman, 2008):

$$P(c_i|c_{i-(n-1)}^{i-1}) \approx \frac{C(c_{i-(n-1)}^{i})}{C(c_{i-(n-1)}^{i-1})} \ , \tag{3}$$

jossa $C(c_{i-(n-1)}^{i})$ tarkoittaa n-grammin $c_{i-(n-1)}^{i}$ esiintymiskertoja opetusaineistossa.

Jos oletetaan eri merkkien todennäköisyydet toisistaan riippumattomiksi, vaidaan kokonaisten tekstilohkojen todennäköisyyksiä edelleen laskea todennäköisyyslaskennan ketjusäännön avulla (Jurafsky ja Martin, 2008, s. 121):

$$P(c_1^i) \approx P(c_1)P(c_2|c_1)\cdots P(c_i|c_{i-n+1}^{n-1}) = \prod_i P(c_i|c_{i-n+1}^{i-1}). \tag{4}$$

Näin saadut arviot (3) ja (4) ovat *suurimman uskottavuuden estimaatteja* (MLE) kyseisille todennäköisyyksille (Jurafsky ja Martin, 2008, s. 122). Niissä n-grammijakaumat (mallin parametrit) on valittu niin, että ne maksimoivat opetusaineiston todennäköisyyden annetuilla n-grammeilla.

Edellä esitelty MLE-kielimalli antaa huonon estimaatin mielivaltaisen n-grammin todennäköisyydelle, jos kyseinen n-grammi on joko harvinainen tai ei esiinny opetusaineistossa lainkaan. Jälkimmäisessä tapauksessa n-grammi saa todennäköisyyden nolla. Tämä johtaa siihen, että arvioitaessa pitkien tekstilohkojen todennäköisyyksiä saadaan todennäköisyydeksi nolla, jos kohdataan yksikin n-grammi, jota ei ole kohdattu opetusaineistossa (vrt. kaava (4)).

Opetusaineistossa harvinaisten n-grammien todennäköisyydet ovat puolestaan karkeasti yliarvioituja. Tämä voidaan ymmärtää seuraavasti: kaikkien mahdollisten n-grammien määrä opetusaineistossa on hyvin suuri, joten kaikkia niitä ei voida tavata edes laajasta korpuksesta. Hyvin harvinaisista eli epätodennäköisistä n-grammeista osa esiintyy opetusaineistossa, osa ei. Niinpä niiden harvinaisten n-grammien todennäköisyysmassa, joita opetusaineistossa ei ole, kertyy niille n-grammeille, jotka aineistossa sattuvat esiintymään. (Virpioja, 2005.)

2.1.1 Tasoitus, perääntyminen ja interpolointi

Harvinaisista ja kohtaamattomista n-grammeista aiheutuvaa ongelmaa pyritään ratkaisemaan kielimallien *tasoituksella*. Tasoituksella tarkoitetaan tekniikoita MLE-todennäköisyyksien korjaamiseksi realistisempaan suuntaan. Termillä viitataan siihen, että tasoitus muokkaa todennäköisyysjakaumaa tasajakauman suuntaan kasvattamalla pieniä ja pienentämällä suuria todennäköisyyksiä eli tasoittaa jakaumaa. (Chen ja Goodman, 1999.)

Yksi ensimmäisistä tasoitusmenetelmistä tunnetaan Laplacen lakina tai yhden lisäyksenä. Siinä kaikkien n-grammien kuvitellaan esiintyneen opetusaineistossa ker-

ran todellista enemmän, jolloin (Manning ja Schütze, 1999, s. 202)

$$P_{\text{Lap}}(c_i|c_{i-1}) = \frac{C(c_{i-1}^i) + 1}{C(c_{i-1}) + |A|} \ , \tag{5}$$

jossa $|A|$ on merkistön koko. Menetelmän on osoitettu antavan liian suuria todennäköisyyksiä opetusaineistossa esiintymättömille n-grammeille. Lidstonen lakina tunnettu tasoitusmenetelmä eroaa Laplacen laista ainoastaan siinä, että n-grammien lukumääriin lisätään ykkösen sijasta jokin pienempi reaaliluku. Lukumäärillä tarkoitetaan tässä opetusaineistosta laskettuja n-grammien esiintymiskertoja. (Manning ja Schütze, 1999, ss. 202–204.)

Good–Turing-tasoitus käyttää kohtaamattomien n-grammien lukumäärän arvioimiseen tietoa kerran nähtyjen n-grammien lukumäärästä. Yleisellä tasolla Good–Turing-tasoituksessa käytetään hyväksi frekvenssien frekvenssejä, jotka kertovat kuinka monta erilaista c kertaa tavattua n-grammia opetusaineistossa on. Olkoon N_c niiden n-grammien määrä, joiden lukumäärä on c. Good–Turing-tasoitus arvioi kunkin luokan lukumäärälle uuden tasoitetun arvion (Jurafsky ja Martin, 2008, s. 135)

$$c^* = (c+1)\frac{N_{c+1}}{N_c} \ . \tag{6}$$

Interpolointi ja *perääntyminen* lähestyvät samaa ongelmaa hieman eri tavalla. Kumpikin menetelmä pyrkii estimoimaan kohtaamattomien n-grammien todennäköisyyksiä lyhyempien n-grammien avulla. Perääntymismenetelmässä siirrytään tutkimaan lyhyempiä n-grammeja, jos pisintä n-grammia ei löydy opetusaineistosta. Esimerkiksi jos opetusaineistosta ei löydy 4-grammia c_{i-4}^i, jotta voitaisiin laskea $P(c_i|c_{i-4}^{i-1})$, peräännytään laskemaan 3-grammijakauman avulla todennäköisyyttä $P(c_i|c_{i-3}^{i-1})$. Perääntyminen voidaan esittää yleisessä muodossa (Chen ja Goodman, 1999)

$$P_{\text{bo}}(c_i|c_{i-n+1}^{i-1}) = \begin{cases} \alpha(c_i|c_{i-n+1}^{i-1}) & \text{jos } C(c_{i-n+1}^i) > 0 \\ \gamma_{c_{i-n+1}^{i-1}} P_{\text{bo}}(c_i|c_{i-n+2}^{i-1}) & \text{muulloin} \end{cases} \ , \tag{7}$$

jossa $\alpha(c_i|c_{i-n+1}^{i-1})$ on tasoitettu ML-estimaatti ja $\gamma_{c_{i-n+1}^{i-1}}$ on skaalaustekijä, jolla saadaan ehdollisten todennäköisyyksien summaksi yksi.

Interpolaatio ottaa huomioon aina myös lyhyemmät n-grammijakaumat, eli interpoloi kaikkien jakaumien kesken (Chen ja Goodman, 1999):

$$P_{\text{interp}}(c_i|c_{i-n+1}^{i-1}) = \alpha(c_i|c_{i-n+1}^{i-1}) + \gamma_{c_{i-n+1}^{i-1}} P_{\text{interp}}(c_i|c_{i-n+2}^{i-1}), \tag{8}$$

jossa $\alpha(c_i|c_{i-n+1}^{i-1})$ ja $\gamma_{c_{i-n+1}^{i-1}}$ ovat jälleen tasoitettu ML-estimaatti ja skaalaustekijä.

Absoluuttinen alennus (Ney ym., 1994) on yksinkertaisempia menetelmiä yhdistävä tasoitusmenetelmä. Menetelmän nimi viittaa siihen, että kaikkien kohdattujen n-grammien lukumäärästä vähennetään vakio alennustermi $D \leq 1$. Näin vapautunut todennäköisyysmassa jaetaan tasan näkemättömien n-grammien kesken (Virpioja, 2005):

$$P_{\text{abs}}(c_i|c_{i-n+1}^{i-1}) = \begin{cases} \frac{C(c_{i-n+1}^i) - D}{C(c_{i-n+1}^{i-1})} & \text{jos } C(c_{i-n+1}^i) > 0 \\ \gamma_{c_{i-n+1}^{i-1}} P_{\text{abs}}(c_i|c_{i-n+2}^{i-1}) & \text{muulloin} \end{cases} \ , \tag{9}$$

jossa $\gamma_{c_{i-n+1}^{i-1}}$ on edelleen skaalaustekijä.

Absoluuttinen alennus perustuu havaintoon, että Good–Turing-tasoitetuissa lukumäärissä c^* (ks. kaava (6)) kunkin luokan lukumäärästä on vähennetty likimain sama vakio (Jurafsky ja Martin, 2008, ss. 143–144). Alkuperäisessä artikkelissaan Ney ym. (1994) ehdottivat alennustermille estimaattia

$$D = \frac{n_1}{n_1 + 2n_2}, \tag{10}$$

jossa n_1 ja n_2 ovat niiden n-grammien määrä, joiden frekvenssi opetusaineistossa on täsmälleen yksi ja kaksi tässä järjestyksessä.

Kneser ja Ney (1995) esittelivät absoluuttisen alennuksen menetelmäänsä täydennyksen, joka on tällä hetkellä yleisesti käytössä oleva tasoitusmenetelmä. Menetelmää kutsutaan *Kneser–Ney-tasoitukseksi* (KN-tasoitus). Menetelmä laajentaa absoluuttista alennusta esittelemällä uuden tavan muodostaa perääntymisessä käytettyjä jakaumia. Sen sijaan että hyödynnettäisiin suoraan lyhyempien n-grammien jakaumia, hyödynnetään erilaisten kontekstien määrää (Virpioja, 2005)

$$N(\bullet c_{i-n+2}^{i}) = |\{c_{i-n+1} : C(c_{i-n+1}^{i}) > 0\}|. \tag{11}$$

Jos erilaisten kontekstien määrä on suuri, voimme olettaa, että tämä merkki esiintyy todennäköisesti myös uusissa konteksteissa. Harvinaiset merkit taas esiintyvät epätodennäköisemmin uusissa konteksteissa. Esimerkkinä voidaan tutkia suomen kielellä opetettua mallia, jonka avulla etsitään todennäköisyyttä $P(\texttt{zsti})$. Opetusdatassa ei ole kohdattu 4-grammia `zsti`, joten peräännytään tutkimaan todennäköisyyttä $P(\texttt{sti})$. Nyt muut menetelmät antavat todennäköisesti aivan liian suuria estimaatteja, koska 3-grammi `sti` on suomenkielessä melko yleinen. Kun sen sijaan tutkitaan erilaisten `z`-kirjaimella alkavien 4-grammien määrää (joka on todennäköisesti todella pieni), huomataan että `z`-kirjaimen esiintyminen uudessa kontekstissa on melko epätodennäköistä. Määritellään

$$N(\bullet c_{i-n+2}^{i-1}\bullet) = \sum_{c_j} N(\bullet c_{i-n+2}^{i-1} c_j) = |\{(c_{i-n+1}, c_i) : C(c_{i-n+1}^{i}) > 0\}|. \tag{12}$$

Tätä merkintää käyttäen KN-tasoitus formalisoidaan seuraavasti (Jurafsky ja Martin, 2008, s. 144):

$$P_{\mathrm{KN}}(c_i|c_{i-n+1}^{i-1}) = \begin{cases} \frac{C(c_{i-n+1}^{i})-D}{C(c_{i-n+1}^{i-1})} & \text{jos } C(c_{i-n+1}^{i}) > 0 \\ \gamma_{c_{i-n+1}^{i-1}} \frac{N(\bullet c_{i-n+2}^{i})}{N(\bullet c_{i-n+2}^{i-1}\bullet)} & \text{muulloin} \end{cases}, \tag{13}$$

jossa $\gamma_{c_{i-n+1}^{i-1}}$ on jälleen sopiva skaalaustekijä. Muunnettu KN-tasoitus (Chen ja Goodman, 1999) eroaa edellisestä siinä, että alennusparametri D optimoidaan erikseen tapauksille, joissa $C(c_{i-n+1}^{i})$ on yksi, kaksi, ja kolme tai enemmän.

2.1.2 Mallien sisäinen arviointi

Tapaa arvioida kielimallin hyvyyttä ilman ulkoista sovelluskohdetta kutsutaan sisäiseksi arvioinniksi (Jurafsky ja Martin, 2008, s. 129). Yksinkertainen tapa arvioida mallin hyvyyttä on sen antama todennäköisyys opetusaineistolle $O = c_1^N$. Varsinaisten todennäköisyyksien sijaan vertaillaan usein *log-todennäköisyyksiä*

$$\log P(O) = \log \left(\prod_{i=1}^{N} P(c_i | c_{i-n+1}^{i-1}) \right) = \sum_{i=1}^{N} \left(\log P(c_i | c_{i-n+1}^{i-1}) \right), \tag{14}$$

joiden laskemisella voidaan välttää pienien todennäköisyyksien kertomisessa uhkaava alivuoto. Log-todennäköisyyksille pätee

$$P(x) < P(y) \quad \Leftrightarrow \quad \log P(x) < \log P(y), \tag{15}$$

joten haluttaessa vertailla mallien antamia todennäköisyyksiä, voidaan yhtä hyvin vertailla vastaavia log-todennäköisyyksiä.

Yleisin sisäisessä arvioinnissa käytetty mitta on *hämmentyneisyys* (Jurafsky ja Martin, 2008, s. 120)

$$PP(O, P) = P(c_1^N)^{-\frac{1}{N}} = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(c_i | c_{i-n+1}^{i-1})}}, \tag{16}$$

joka riippuu sekä kielimallista P että opetusaineistosta O (kielestä). Hämmentyneisyys voidaan tulkita mallin antamaksi kielen haarautumistekijän geometriseksi keskiarvoksi (Rosenfeld, 2000). Haarautumistekijällä tarkoitetaan tässä mitä tahansa merkkiä seuraavien mahdollisten merkkien määrää.

2.1.3 Karsiminen

Opetusaineistoa tai n-grammien pituutta kasvatettaessa myös kielimallin koko kasvaa nopeasti. Tämän vuoksi on tavallista rajoittaa kielimallin sisältämien n-grammien määrää. Yleinen lähestymistapa on rakentaa opetusaineistosta täysi malli haluttuun pituuteen saakka, ja sen jälkeen *karsia* mallista n-grammeja jonkin säännön mukaan. (Siivola ym., 2007)

Kneser (1996) on esitellyt yleisen karsimismenetelmän perääntymismalleille (*Kneser pruning*, KP), joka laskee painotetun muutoksen log-todennäköisyydessä poistettaessa n-grammi mallista. KP-menetelmä käyttää log-todennäköisyyden muutoksen laskemiseen absoluuttisen alennuksen mallia ja muutos painotetaan mallin antamalla todennäköisyydellä.

KP-menetelmä ei ota huomioon KN-tasoituksen ominaisuuksia. Siivola ym. (2007) ehdottavat Kneserin menetelmään korjausta (*revised Kneser pruning*), joka ottaa KN-tasoituksen ominaisuudet paremmin huomioon. Korjattu menetelmä päivittää KN-tasoitetun mallin perääntymisjakaumat karsittavan n-grammin poiston jälkeen

ja laskee muutoksen opetusaineistolle lasketussa log-todennäköisyydessä vanhalla ja päivitetyllä mallilla. Jos muutos on suurempi kuin karsimisparametri ϵ, palautetaan n-grammi malliin. Muulloin poistaminen on järkevää ja se jää voimaan.

2.1.4 N-grammimalli luokittimena

Kielentunnistuksessa tehtävänä on luokitella mielivaltainen tekstilohko C johonkin kieleen K, jota kuvaa kielimalli P_K, eli maksimoida todennäköisyyttä $P(K|C)$. Sovellamme Bayesin teoreemaa:

$$P(K|C) = \frac{P(C|K)P(K)}{P(C)} \ = P(C|K) = P_K(C), \tag{17}$$

jossa $P(C) = 1$, eri kielten prioritodennäköisyydet $P(K)$ ovat yhtä suuria ja K kuuluu kaikkien opetusaineistoista saatavien kielimallien joukkoon. Kun etsimme sen kielimallin P_K, jolla todennäköisyys $P_K(C)$ maksimoituu, saamme *suurimman todennäköisyyden luokittimen*

$$K_{\mathrm{ML}} = \operatorname*{argmax}_K P_K(C). \tag{18}$$

2.2 Informaatioteoriaa

Olkoon X diskreetti satunnaismuuttuja joukossa $\mathcal{X}$ ja $P(X)$ X:n todennäköisyysjakauma. Satunnaismuuttujan X *entropia* (*entropy*) määritellään (Shannon, 1948)

$$H(X) = -\sum_{x \in \mathcal{X}} P(x) \log P(x). \tag{19}$$

Entropia mittaa satunnaismuuttujan informaation määrää. Yleisimmin käytetään kaksikantaista logaritmia, jolloin entropia mittaa informaatiota bitteinä. Intuitiivisen tulkinnan mukaan entropia on pienin tarvittava määrä bittejä koodaamaan satunnaismuuttujan arvo optimaalisella koodauksella. (Jurafsky ja Martin, 2008, s. 148)

Liittääksemme entropian tilastollisen kielimallinnuksen viitekehykseen, tarvitsemme vielä ristientropian käsitettä. Kieli voidaan tulkita todennäköisyysjakauman $P(X)$ mukaan jakautuneeksi diskreetiksi satunnaismuuttujaksi X kaikkien mahdollisten merkkijonojen joukossa $\mathcal{X}$. Tällöin kielimallin P_K ja todennäköisyysjakauman $P(X)$ (kielen) välinen ristientropia on (Rosenfeld, 2000)

$$H_{\mathrm{C}}(P, P_K) = -\sum_{x \in \mathcal{X}} P(x) \log P_K(x). \tag{20}$$

Ristientropia riippuu useasta tekijästä: kielimallista, kielestä, arvioitavasta tekstistä ja yksiköistä, joihin malli perustuu (esim. merkit, morfeemit tai sanat) (Virpioja, 2005).

Kielimallien arvioinnissa uudella aineistolla käytetään usein *keskimääräistä log-todennäköisyyttä* (Rosenfeld, 2000)

$$\text{Average-Log-Likelihood}(C|P_K) = \frac{1}{N} \log P_K(C) = \frac{1}{N} \sum_{i=1}^{N} \log P_K(c_i), \qquad (21)$$

jossa $C = c_1^N$ ja P_K on annettu kielimalli. Jotta jälkimmäinen yhtäsuuruus pitäisi paikkaansa, täytyy merkkien todennäköisyydet olettaa toisistaan riippumattomiksi. Keskimääräinen log-todennäköisyyden vastaluku voidaan tulkita myös merkkien määrällä normalisoiduksi empiiriseksi ristientropiaksi tekstilohkolle C ja kielimallille P_K:

$$H_{\mathrm{C}}(C, P_K) = -\frac{1}{N} \sum_{i=1}^{N} \log P_K(c_i) = -\frac{1}{N} \log P_K(C). \qquad (22)$$

Ristientropiaa voidaan tulkita kuten entropiaakin: ristientropia $H_C(C, P_K)$ mittaa, kuinka monta bittiä merkkiä kohden tarvitaan viestin C koodaamiseen koodauksella P_K.

Edellä esitetty tulkinta liittää kielimalleihin pohjautuvan kielentunnistimen ristientropiaa vertaileviin menetelmiin (Sibun ja Reynar, 1996; Teahan, 2000). Teahan ja Harper (2001) ovat kehittäneet menetelmäänsä tekstin pakkauksessa yleisesti käytettyn PPM-pakkausmenetelmän (Cleary ym., 1984) pohjalta. Sibun ja Reynar (1996) käyttävät n-grammimalleja, jotka on tasoitettu eräänlaisella *Lidstonen lain* muunnoksella. Malleissa jokainen n-grammi, joka on nähty muissa kuin mallinnettavassa kielissä, katsotaan nähdyksi 0,5 kertaa.

2.3 Vertailumenetelmät

Jos halutaan vertailla samaa ongelmaa eri tavalla lähestyviä menetelmiä tasapuolisesti, tarvitaan sekä testausaineistoa johdonmukaisesti hyödyntävä validointimenetelmä että tilastollisia menetelmiä tulosten merkittävyyden analysointiin. Tässä alaluvussa esitellään aluksi erityisesti pienille testiaineistoille soveltuva ristiinvalidointimenetelmä. Sen jälkeen esitellään tulosten merkitsevyyttä testaavat tilastolliset testit.

Kielimallien antamat todennäköisyydet estimoidaan opetuksessa käytettävästä kieliaineistosta. Jotta mallien hyvyyttä voitaisiin mitata, täytyy niitä testata aineistolla, jota malli ei ole nähnyt. Opetusaineistoa käyttävillä menetelmillä on yleinen taipumus *ylioppia* aineisto, mikä tarkoittaa sitä, että malli ennustaa tulevat tapahtumat liikaa opetusaineiston kaltaisiksi. Siksi on oleellista opettaa malli ja testata sen toimivuutta eri aineistolla. (Manning ja Schütze, 1999, ss. 206–207.)

Kielentunnistukseen käytettävien luokittimien vertailemiseksi jaetaan kieliaineiston opetus- ja testausosiin. Opetusaineistoa käytetään luokittimen n-grammimallien opettamiseen (ks. kuva 2 sivulla 13), jonka jälkeen luokittimen suoriutumista testataan testiaineistosta otetuilla eri mittaisilla näytteillä. Tämän lisäksi tarvitaan

usein opetusaineistosta riippumaton *validointiaineisto*, jolla voidaan estimoida mallin opetuksessa käytettäviä parametreja. (Jurafsky ja Martin, 2008, ss. 125–126.)

2.3.1 Ristiinvalidointi

Ristiinvalidoinnissa pyritään muodostamaan testimateriaalista $\mathcal{X}$ opetus- ja testausaineistopareja $\{\mathcal{O}_i, \mathcal{T}_i\}_i$, joilla voidaan arvioida käytettyä menetelmää. Ristiinvalidointi käyttää koko testimateriaalia jokaisella validointikierroksella välttäen samalla opetus- ja testausaineistojen päällekkäisyyden, joten sen antamat virhe-estimaatit ovat mahdollisimman tarkkoja. Erityisesti kaikki osat testimateriaalista pääsevät vaikuttamaan sekä mallin opetukseen että sen testaukseen. Aineiston tarkka hyödyntäminen on tärkeää varsinkin silloin kun se on suppea. (Alpaydin, 2004, ss. 330–331.)

K-kertaisessa ristiinvalidoinnissa testimateriaali $\mathcal{X}$ – tässä tapauksessa tekstiaineisto – jaetaan K:hon yhtä suureen osaan $\{\mathcal{X}_i \mid i = 1, \ldots, K\}$. Testausparit muodostetaan niin että jokainen osa $\mathcal{X}_i$ toimii vuorollaan testausaineistona, ja muut osat yhdistetään opetusmateriaaliksi:

$$\left.\begin{array}{l} \mathcal{T}_i = \mathcal{X}_i \\ \mathcal{O}_i = \mathcal{X} \setminus \mathcal{X}_i \end{array}\right\} \qquad i = 1, \ldots, K. \tag{23}$$

2.3.2 K-kertaisesti validoitu t-testi

Kahden eri luokittimen suoriutumista voidaan verrata *K-kertaisesti ristiinvalidoidulla t-testillä*. Testi paljastaa, suorittavatko luokittimet tehtävänsä yhtä hyvin tilastollisella merkitsevyystasolla α. (Alpaydin, 2004, ss. 342–343)

Olkoot p_i^1 ja p_i^2 luokittimien virheprosentit ristiinvalidoinnin kierroksella i ja virheprosenttien erotus $p_i = p_i^1 - p_i^2$. Määritellään erotusten keskiarvo m ja otosvarianssi S^2:

$$m = \frac{\sum_{i=1}^{K} p_i}{K} \ , \quad S^2 = \frac{\sum_{i=1}^{K} (p_i - m)^2}{K - 1} \ . \tag{24}$$

K-kertaisessa ristiinvalidoinnissa saamme K kappaletta arvoja p_i, jotka ovat (likimain) normaalisti jakautuneita. Testin nollahypoteesi on, että tämän jakauman keskiarvo μ on 0:

$$\begin{array}{l} H_0 : \mu = 0 \\ H_1 : \mu \neq 0 \end{array} \tag{25}$$

Määritellään t-testisuure, joka noudattaa nollahypoteesin ollessa voimassa Studentin t-jakaumaa vapausastein $(K - 1)$:

$$t = \frac{\sqrt{K} \cdot m}{S} \sim t_{K-1}. \tag{26}$$

Nollahypoteesi hyväksytään merkitsevyystasolla α, jos t-testisuureen arvo on välillä $(-t_{\alpha/2,K-1}, t_{\alpha/2,K-1})$. Merkitsevyystaso α on todennäköisyys sille, että nollahypoteesi hylätään virheellisesti (*1. lajin virhe*). (Alpaydin, 2004, ss. 342–343)

Sen sijaan että annettaisiin merkitsevyystasolle etukäteen jokin kiinteä arvo, on myös mahdollista etsiä pienin merkitsevyystaso α, jolla nollahypoteesi voidaan hylätä. Tätä merkitsevyystasoa kutsutaan testin *p-arvoksi*. Nollahypoteesi hylätään, jos p-arvoa voidaan pitää riittävän pienenä. (Milton ja Arnold, 2002, ss. 273–275.)

2.3.3 Fisherin menetelmä

Fisher (1948) on kehittänyt menetelmän riippumattomien p-arvojen yhdistämiseksi. Menetelmässä testataan samaa nollahypoteesia kuin p-arvoja laskettaessa. Määritellään X^2 testisuure, joka noudattaa nollahypoteesin ollessa voimassa χ^2-jakaumaa vapausastein $2k$:

$$X^2 = -2\sum_{i=1}^{k} \log p_i \sim \chi^2_{2k}, \tag{27}$$

jossa k on yhdistettävien p-arvojen lukumäärä ja p_i on i. testin p-arvo. Nollahypoteesi hyväksytään jos testisuuren arvo on välillä $(\chi^2_{\alpha/2,2k}, \chi^2_{1-\alpha/2,2k})$. Vaihtoehtoisesti voidaan jälleen määrittää testin merkitsevyystaso, eli p-arvo.

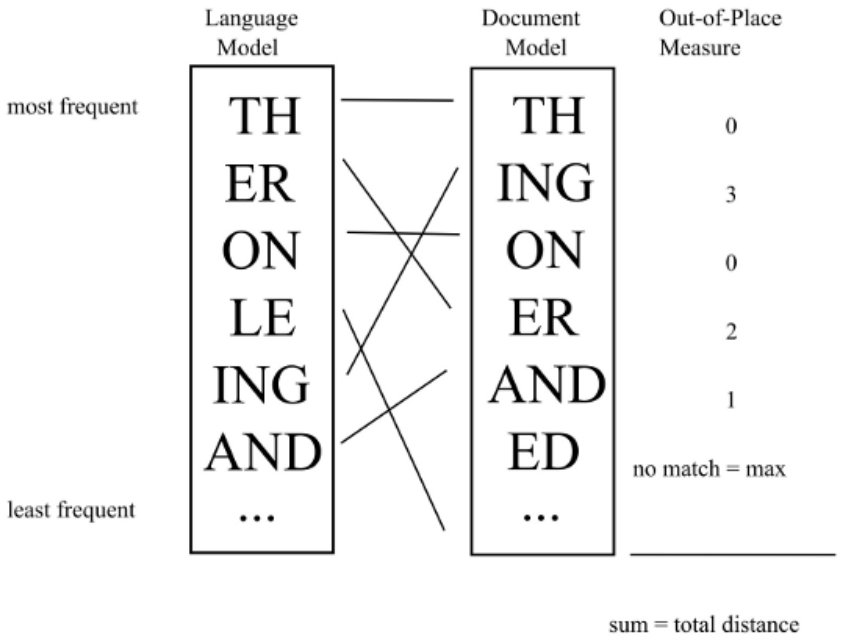
2.4 Kielentunnistusmenetelmät

Aikaisemmasta kirjallisuudesta voidaan löytää kirjava joukko erilaisiin käyttötarkoituksiin kehitettyjä kielentunnistusmenetelmiä. Tuomainen (2008) on tehnyt katsauksen muutamaan yleiseen tilastolliseen menetelmään. Näitä ovat Cavnarin ja Trenklen (1994) yleisiä n-grammeja käyttävä menetelmä, Dunningin (1994) tilastollinen menetelmä, Teahanin ja Harperin (2001) PPM-pohjainen menetelmä ja Poutsman (2001) Monte Carlo -menetelmä.

Botha ja Barnard (2007) listaavat kielentunnistamisen tarkkuuteen vaikuttavia tekijöitä tutkivassa artikkelissaan suuren joukon kielentunnistuksessa käytettyjä menetelmiä. Heidän artikkelista nostettakoon esiin, että (i) Kruengkrai ym. (2005) käyttävät kielen tunnistamiseen merkkijonoytimiä ja tukivektorikoneita, (ii) Damashek (1995) soveltaa menetelmässään normalisoitua pistetuloa ja (iii) Sibun ja Reynar (1996) käyttävät suhteellista entropiaa. Lisäksi Botha ja Barnard (2007) mainitsevat käytetyn päätöspuita, neuroverkkoja ja usean muuttujan lineaarista regressiota.

2.4.1 Ranking-menetelmä

Cavnar ja Trenkle (1994) ovat kehittäneet Ranking-menetelmän, joka on edelleen erittäin käytetty menetelmä sekä kielen tunnistamiseen että yleisempään tekstien kategorisointiin. Tekstien kategorisoinnilla tarkoitetaan niiden luokittelua esimerkiksi kirjoittajan tai genren mukaan. Ranking-menetelmässä jokaisesta kielestä luodaan profiili, joka sisältää opetusmateriaalin K yleisintä n-grammia ($n = 1, \ldots, N$) järjestyksessä yleisimmästä harvinaisimpaan. Kokeellisessa osuudessa käytettiin yleisestä käytettyjä opetusparametrejä $K = 400$ ja $N = 4$.

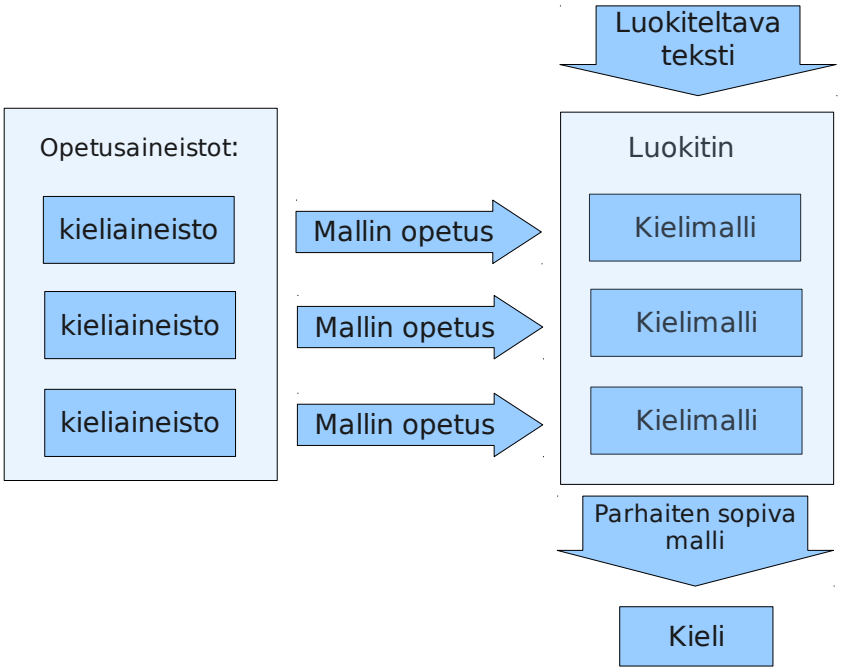


Kuva 1: Esimerkki Ranking-menetelmän toiminnasta. Kuvassa esitetyt profiilit ovat menetelmän selkeyttämiseksi keksittyjä, eivätkä mukaile kieliaineistosta laskettujen kieliprofiilien rakennetta. (Cavnar ja Trenkle, 1994.)

Tekstin luokittelemiseksi myös siitä luodaan edellä kuvatun kaltainen profiili (jatkossa testiprofiili), jota verrataan eri kielten (tai kategorioiden) profiileihin. Profiilien välisen etäisyyden laskeminen on esitetty kuvassa 1. Kuvassa olevat englanninkieliset termit on jatkossa esitetty suluissa. Testiprofiilille (Document model) lasketaan etäisyys (Out-of-Place Measure) kuhunkin profiiliin (Category Profile) summaamalla kaikkien n-grammien etäisyydet (sum = total distance). Tämän jälkeen teksti luokitellaan siihen profiiliin, johon sen etäisyys on pienin. Profiileissa yleimmät (most frequent) n-grammit on kuvattu ylös ja vähiten yleiset (least frequent) alas. Jos n-grammia ei löydy vertailtavasta profiilista (no-match) annetaan sille jokin ennalta määrätty maksimiarvo (max).

2.4.2 Dunningin menetelmä

Dunning (1994) oli yksi ensimmäisistä n-grammimalleja kielentunnistamiseen soveltaneista tutkijoista. Paljon viitatussa artikkelissaan hän mallinsi kieliä Laplacen lain mukaan tasoitetuilla n-grammimalleilla ja sovelsi suurimman todennäköisyyden luokitinta siten, että tekstilohko tulkitaan kuuluvaksi siihen kieleen, jonka kielimalli antaa tekstilohkolle suurimman todennäköisyyden. Tällaisen, kielimalleihin perustuvan luokittimen toiminta on esitetty kuvassa 2. Dunningin menetelmässä kielimallien opetus tapahtuu kaavan (5) mukaisesti. Todennäköisyyksien laskemiseksi tarvittava



Kuva 2: Kielimallipohjaisen kielentunnistimen toiminta. Diagrammilla voidaan kuvata sekä Dunningin LLM-menetelmän että tässä työssä esitellyn KNLM-menetelmän toimintaa.

merkistön koko laskettiin jokaisen kielen opetusaineistosta niin, että kullakin mallilla oli oma merkistön koko. Menetelmästä käytetään jatkossa nimeä LLM (Laplace Language Model).

Dunning testasi luokitintaan eri n:n arvoilla ja osoitti, että tunnistustarkkuus huononee jos käytetään arvoja $n > 3$. Opetusaineiston ollessa pieni (2 kilotavua) parhaat tulokset syntyivät arvolla $n = 2$. Dunning testasi menetelmäänsä myös 50 kilotavun opetusaineistoilla, jolloin n:n arvot 2 ja 3 olivat tuottivat likimain samanlaisia tuloksia. Tämän työn kokeellisessa osuudessa käytettyjen opetusaineistojen koot olivat Dunningin käyttämien aineistokokojen välissä. Testeissä havaittiin, että parhaat tulokset syntyivät kun $n = 2$.

2.4.3 KNLM-menetelmä

Tässä työssä esitellään kielentunnistusmenetelmä, jota kutsumme KNLM (Kneser-Ney Language Model) menetelmäksi. Menetelmä mallintaa kieliä Kneser–Ney-tasoitetuilla ja karsituilla kielimalleilla, jotka on luotu Siivolan ym. (2007) implementoimalla VariKN työkalulla[2]. Kielimallien opetuksessa käytettiin parametria $n = 4$. Menetelmän toiminta on esitetty kuvassa 2. Erot KNLM- ja LLM-menetelmien vä-

[2]Saatavilla osoitteessa `http://varikn.forge.pascal-network.org/`.

lillä ovat kielimallien opetuksessa. KNLM-luokittimen kielimalleissa käytetään KN-tasoitusta (kaava (13)) ja kielimallista karsitaan piirteitä Siivolan ym. (2007) korjatulla KP-algoritmilla. Annettu tekstilohko luokitellaan kuuluvaksi siihen kieleen, jonka kielimalli antaa tekstilohkolle suurimman todennäköisyyden.

3 Kokeellinen osuus

Kokeellisen osuuden tarkoituksena oli tutkia kuinka n-grammimalleihin pohjautuva kielentunnistusmenetelmä suoriutuu lyhyiden tekstilohkojen tunnistamisesta. Kokeissa vertailtiin tasoitettuihin ja karsittuihin n-grammimalleihin pohjautuvaa luokitinta Dunningin (1994) sekä Cavnarin ja Trenklen (1994) ehdottamiin menetelmiin. Vertailu Dunningin LLM-menetelmään on erityisen mielenkiintoinen, koska artikkelissaan Dunning (1994) toteaa, ettei kielentunnistuksessa käytettävissä n-grammimalleissa ole tarpeellista käyttää edistyneempiä tasoitusmenetelmiä.

3.1 Tutkimusaineisto

Tutkittaessa kielentunnistusmenetelmän toimintaa tarvitaan monikielinen kieliaineisto. Jotta testitulokset olisivat vertailukelpoisia, tulisi aineiston sisältää sama määrä saman tyylistä materiaalia kaikilla kielillä. Ideaalinen monikielinen aineisto sisältää saman tekstin käännettynä kaikille kielille samaa tarkoittavat fraasit kohdistettuna, mikä on olennaista konekäännösmallien luomisessa.

Tässä tutkimuksessa käytettiin kieliaineistona YK:n ihmisoikeusjulistusta, joka on käännetty 364:lle kielelle.[3] Koska tekstin irrottaminen pdf-tiedostoista käsiteltävään muotoon ei ollut mahdollista kaikista dokumenteista, oli kokeissa käytössä enimmillään 298 kieltä. Tekstin irrottamisessa käytettiin Linuxin `pdftotext`-työkalua ja tekstit tallennettiin käyttäen UTF-8 merkistökoodausta, joka pystyi esittämään kaikkien käytössä olleiden kielten merkistöt. Aineistosta poistettiin ylimääräiset välilyönnit ja rivinvaihdot.

Ihmisoikeusjulistuksen pituudessa oli suurta kielikohtaista vaihtelua. Taulukossa 1 on listattu aineistojen koko valikoiduilla kielillä. Merkeissä mitattuna aineiston koon mediaani oli 11201 merkkiä ja kvartiilivälin rajat 10093 ja 12537 merkkiä. Kvartiiliväli on se 50 % osuus aineistosta, jonka molemmille puolille jää 25 % osuus aineistoa. Kaikkein poikkeuksellisimmat arvot on lihavoitu taulukkoon 1. Suppeimpien aineistojen kohdalla on todennäköistä, että myös tekstin sisältöä on karsittu. Mitattaessa aineistoja tiedoston koolla saadaan mediaaniksi 11,6 kilotavua ja kvartiilivälin rajoiksi 10,2 ja 13,1 kilotavua. Kielikohtaisten merkistöjen kokojen mediaani oli 61 merkkiä ja kvartiilivälin rajat 57 ja 68 merkkiä. Kieliä, joiden merkistön koko ylitti 100 merkkiä, oli seitsemän. Suurimmat merkistöt olivat kiinan (539 merkkiä) ja japanin (506 merkkiä) kielillä.

Kokeissa käytettiin kahta erillistä kielijoukkoa. Täysi 298 kielen joukko sisälsi kaikki kielet, jotka pystyttiin irrottamaan ihmisoikeusjulistuksesta käsiteltävään muotoon. Lisäksi käytettiin 60 kielen joukkoa, joka koostui sekä ihmisoikeusjulistuksen että Google AJAX language APIn[4] kielivalikoimista löytyvistä kielistä.

[3] YK:n ihmisoikeusjulistus on saatavilla osoitteessa `http://www.ohchr.org/EN/UDHR/Pages/Introduction.aspx`. Käännösmäärä on tarkistettu 27.10.2009.

[4] Saatavilla osoitteessa `http://code.google.com/apis/ajaxlanguage/`.

Taulukko 1: Kieliaineistojen ja kielikohtaisten merkistöjen koot valikoiduilla kielillä. Ääriarvot on lihavoitu ja keskiarvot on laskettu kaikista 298 kielestä.

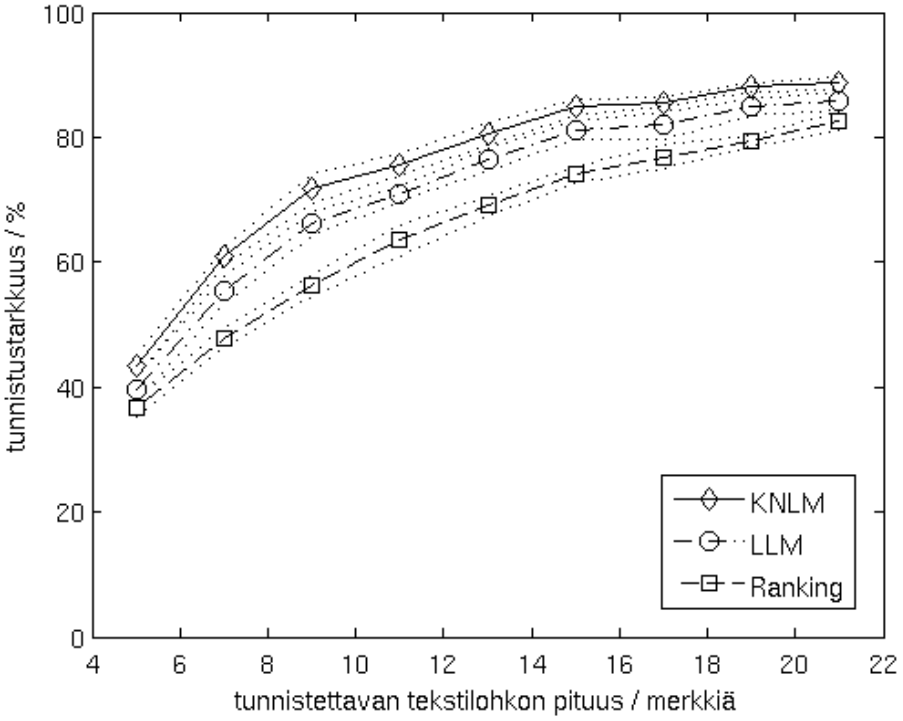
Kieli	**Opetusaineiston koko** (kilotavua)	(merkkiä)	**Merkistön koko** (merkkiä)
suomi	12,6	12328	57
englanti	10,5	10730	57
saksa	12,0	12065	70
ranska	12,1	11990	60
espanja	12,0	12060	61
tšekki	11,0	9912	79
kreikka	22,3	12529	68
venäjä	21,3	11902	67
tagalog	13,2	13488	57
divedin kieli	**41,2**	21731	69
cashinahua	**5,0**	5016	59
asháninca	26,4	**26844**	60
japani	12,2	**4368**	506
kiina	10,9	5457	**539**
kikongo	11,7	11931	**44**
Keskiarvo	12,2	11467	67

3.2 Validointi

Kokeissa käytetty kieliaineisto oli suhteellisen pieni, joten oli perusteltua käyttää ristiinvalidointia. Koska kielentunnistimien toimintaa haluttiin testata erityisesti lyhyillä merkkijonoilla, valittiin validoinnissa käytettävät näytteet satunnaisesti kustakin testausaineistosta. Kaikkea testausaineistoa ei siis todellisuudessa käytetty mallien testaamiseen, mutta tätä ei pidä tulkita niin, etteikö koko aineistoa olisi käytetetty vuorollaan opetusaineistona. Kokeissa käytettiin siis hieman muunnettua 10-kertaista ristiinvalidointia, koska ristiinvalidointi kaikella materiaalilla olisi ollut todella aikaavievä prosessi. Koska ihmisoikeusjulistuksen pituudessa oli suurta kielikohtaista vaihtelua, vaihtelivat myös opetusaineistojen koot huomattavasti eri kielien välillä.

3.3 Kokeet 298 kielellä

Kielentunnistimia vertailtiin tekemällä jokaiselle luokittimelle 10-kertainen ristiinvalidointi testimateriaalina toimineella YK:n ihmisoikeusjulistuksella, joka oli käytettävissä 298 kielellä. Jokaiselta ristiinvalidoinnin kierrokselta laskettiin kullekkin



Kuva 3: KNLM-, LLM- ja Ranking-menetelmien vertailu 298 kielellä.

tekstilohkon pituudelle keskimääräinen tunnistustarkkuus, ja lopullinen tarkkuus oli näiden kymmenen arvon keskiarvo. Luokittimien suoriutuminen on nähtävissä kuvassa 3, jossa pisteviivat osoittavat tulosten 95 % luottamusvälin. Kuvan perusteella KNLM luokitin suoriutuu kielen tunnistamisesta vertailumenetelmiä paremmin. Kuvasta voidaan myös nähdä, että LLM-menetelmä suoriutuu Ranking-menetelmää paremmin, ja että KNLM- ja LLM-menetelmien erot tasoittuvat testitekstien pituuden kasvaessa.

Luokittimien suoriutuminen valikoiduilla kielillä on nähtävissä taulukossa 2. Tarkkuudet kuvaavat sitä, kuinka monta prosenttia eri kielellä olevista näytteistä kukin menetelmä tunnisti oikein. Taulukosta voidaan nähdä, että vaikka KNLM luokitin suoriutuu keskimääräisesti muita menetelmiä paremmin, tunnistaa se joitain kieliä, kuten englannin, vertailumenetelmiä huonommin. LLM-menetelmä puolestaan tunnistaa huonosti suuren merkistön omaavia kieliä, kuten kiina. Englannin- ja espanjankielisten näytteiden tunnistaminen osoittautui hankalaksi kaikille kielentunnitimille. Syitä edellä esitettyihin poikkeamiin on pohdittu luvussa 4.

3.4 Kokeet 60 kielellä

Kielentunnistimia vertailtiin myös käyttäen 60 kielen joukkoa, jotta tuloksia pystyttiin vertailemaan Google AJAX language APIn kanssa. Opetusaineistona KNLM-,

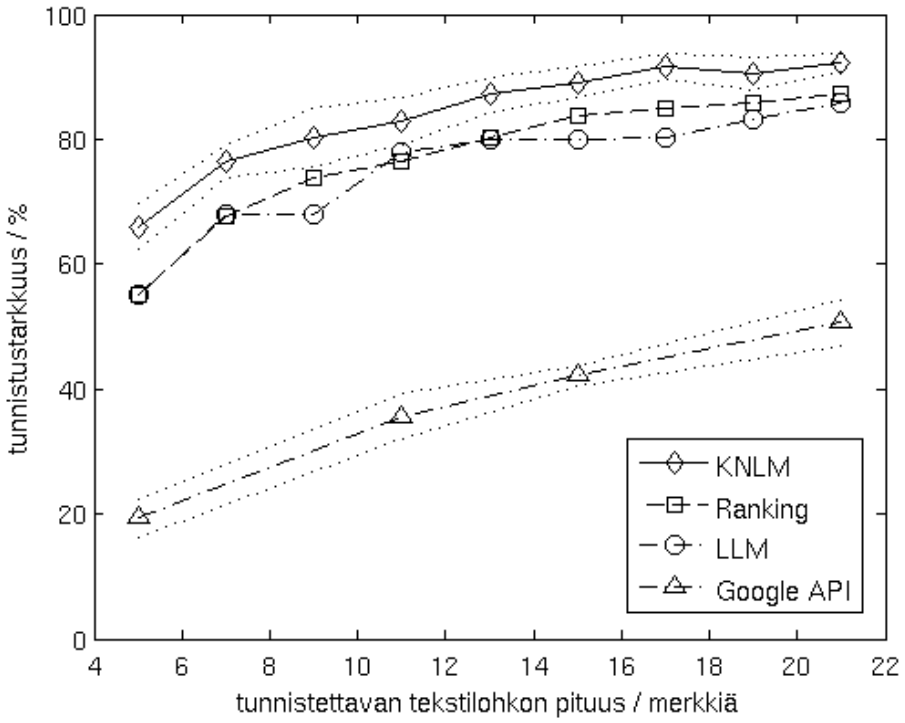
Taulukko 2: Kielentunnistustarkkuudet prosentteina valikoiduilla kielillä olleista näytteistä sekä kaikkien kielien keskiarvot 298 kielen joukosta käyttäen KNLM-, LLM- ja Ranking-menetelmiä.

	KNLM			**LLM**			**Ranking**		
	näytteen pituus (merkkiä)			näytteen pituus (merkkiä)			näytteen pituus (merkkiä)		
Kieli	5	11	21	5	11	21	5	11	21
englanti	17,3	24,9	23,2	19,6	39,5	52,7	12,7	38,0	66,5
suomi	40,0	84,3	98,9	45,5	84,6	97,9	36,0	64,5	92,0
saksa	41,6	80,0	95,1	33,6	70,1	91,5	26,9	63,3	81,8
ranska	23,7	60,6	85,6	23,7	60,1	86,6	17,8	41,1	68,0
espanja	8,1	35,1	59,0	6,4	31,3	54,6	2,4	17,1	38,9
tšekki	36,5	65,9	80,5	26,1	53,9	74,7	30,9	54,7	75,8
kreikka	99,7	100,0	100,0	86,5	97,0	99,2	97,8	99,8	100,0
venäjä	51,9	78,3	89,4	64,0	86,0	93,9	44,2	63,3	78,9
kiina	92,4	98,1	98,9	1,8	2,9	0,2	83,8	98,7	100,0
uzbekki	40,4	86,6	98,6	34,9	76,6	94,0	30,4	58,8	89,4
tagalog	21,2	57,6	82,4	15,8	52,5	78,5	18,6	43,2	71,2
malaiji	16,8	46,2	52,4	16,9	41,1	47,9	16,2	30,8	44,4
keskiarvo	43,3	75,6	88,6	39,7	70,9	85,7	36,7	63.5	82,6

LLM- ja Ranking-menetelmille käytettiin edelleen YK:n ihmisoikeusjulistusta. Jokaiselle luokittimelle tehtiin 10-kertainen ristiinvalidointi, jossa jokaiselta ristiinvalidoinnin kierrokselta laskettiin kullekin tekstilohkon pituudelle keskimääräinen tunnistustarkkuus. Googlen APIn testaamisessa päätettiin käyttää vain neljää eri testisyötteen pituutta, jotka olivat 5, 11, 15 ja 21 merkkiä, jotta vältyttiin liialliselta palvelun kuormittamiselta.

Luokittimien suoriutuminen on nähtävissä kuvassa 4. Pisteviivat, jotka piirrettiin kuvan selkeyttämiseksi ainoastaan KNLM-menetelmälle ja Google APIlle, osoittavat tulosten 95 % luottamusvälin. LLM- ja Ranking-menetelmien tulosten luottamusvälit olivat samaa suuruusluokkaa. Kuvasta on nähtävissä, että KNLM-luokitin suoriutuu kielen tunnistamisesta jälleen vertailumenetelmiä paremmin. Ranking-menetelmä suoriutui tällä aineistolla likimain samalla tavoin LLM-menetelmän kanssa.

Kaikkien neljän kielentunnistimen suoriutuminen valikoiduilla kielillä on nähtävissä taulukossa 3. Tarkkuudet kuvaavat sitä, kuinka monta prosenttia eri kielellä olevista näytteistä kukin menetelmä tunnisti oikein. Taulukosta voidaan nähdä, että parhaiten suoriutuivat KNLM- ja LLM-menetelmät, joiden merkittävimmät suorituserot löytyvät kiinan kielen ja joidenkin 5 merkin pituisten näytteiden tunnistamisesta.



Kuva 4: KNLM-, LLM- ja Ranking-menetelmien ja Google APIn vertailu 60 kielellä.

Ranking-menetelmä osoitti useilla kielillä KNLM- ja LLM-menetelmien tasoista tunnistustarkkuutta, mutta esimerkiksi englannin-, ranskan- ja venäjänkielisten näytteiden kohdalla se ei yltänyt kielimallipohjaisten luokittimien tasolle. Englannin- ja espanjankielisten näytteiden tunnistaminen osoittautui hankalaksi kaikille luokittimille myös tässä kielijoukossa.

Google API tunnisti Euroopan valtakielet englannin, saksan ja ranskan erittäin luotettavasti varsinkin pisimmistä, 21 merkin mittaisista näytteistä. Harvinaisempina kielinä taulukkoon mukaan otettujen uzbekin ja malaijin kielten tunnistamisessa Googlella oli hankaluuksia ja tagalin kieltä Google ei tunnistanut lainkaan, vaikka se löytyy heidän tunnistuslistaltaan.[5] Myöskään 21 merkin mittaisia kreikankielisiä näytteitä Google ei tunnistanut. Yleisesti ottaen taulukosta 3 voidaan nähdä, että poikkeuksia lukuunottamatta Googlen tulokset parantuivat näytteiden pituuden kasvaessa muita menetelmiä enemmän. Luvussa 4 on pohdittu myös edellä esitettyjen poikkeamien syitä.

[5]Tagalog eli tagalin kieli on yksi tärkeimmistä Filippiineillä puhutuista kielistä ja sitä puhuu noin 22 miljoonaa ihmistä.

Taulukko 3: Kielentunnistustarkkuudet prosentteina valikoiduilla kielillä olleista näytteistä 60 kielen joukosta sekä kaikkien kielien keskiarvot käyttäen KNLM-, LLM- ja Ranking-menetelmiä sekä Google AJAX language APIa.

	KNLM			**LLM**			**Ranking**			**Google API**		
	näytteen pituus (merkkiä)			näytteen pituus (merkkiä)			näytteen pituus (merkkiä)			näytteen pituus (merkkiä)		
Kieli	5	11	21	5	11	21	5	11	21	5	11	21
englanti	44,8	77,4	93,4	48,6	73,1	93,8	36,0	64,4	89,8	63,5	88,5	99,5
suomi	59,6	89,6	98,6	62,5	91,2	98,9	48,2	85,0	96,8	18,0	46,5	92,0
saksa	53,4	85,6	95,8	47,7	76,7	92,2	39,2	69,8	87,8	24,5	59,5	96,5
ranska	41,4	72,6	93,8	46,3	75,5	94,8	37,4	62,8	80,6	33,0	76,0	99,0
espanja	26,8	51,4	68,0	26,7	49,7	66,2	16,4	39,4	66,4	17,0	57,0	92,5
tšekki	45,0	65,8	80,0	34,9	60,1	76,7	35,2	56,0	77,0	27,0	65,5	91,0
kreikka	99,8	100	100	91,8	99,4	99,5	98,0	100	99,8	100	100	0,0
venäjä	58,2	77,0	90,6	66,2	86,2	95,1	49,2	70,2	82,0	45,5	83,0	82,5
kiina	94,6	97,6	99,0	4,4	3,5	1,7	87,0	98,4	99,4	3,0	11,0	100
uzbekki	64,8	93,2	98,8	59,5	86,2	97,2	51,4	74,8	94,0	4,0	2,0	21,0
tagalog	69,6	91,6	99,2	71,9	92,7	98,3	62,8	87,4	96,6	0,0	0,0	0,0
malaiji	35,4	53,2	57,6	30,4	44,8	50,7	28,8	44,2	53,2	7,0	30,0	42,0
Keskiarvo	66,0	82,8	92,2	55,0	77,8	85,8	55,0	76,3	87,3	19,3	35,7	50,7

3.5 Tilastollinen testaus

Eri menetelmien suoriutumisen välillä olevien erojen tilastollista merkitsevyyttä mitattiin parittaisella t-testillä. Nollahypoteesina oli, että menetelmien keskimääräinen suoriutuminen eri ristiinvalidoinnin kierroksilla on yhtä hyvää. T-testiä ja Fisherin menetelmää voidaan käyttää, koska eri testitulokset ovat toisistaan riippumattomia.

KNLM- ja LLM-menetelmien suoriutumista 298 kielen joukossa vertailtiin laskemalla p-arvot kullekkin tekstipituudelle parittaisella t-testillä. Arvot vaihtelivat välillä $7,4 \times 10^{-5}$–$5,1 \times 10^{-3}$, minkä perusteella voidaan suoraan todeta KNLM-menetelmän suoriutuvan tilastollisesti merkitsevästi LLM-menetelmää paremmin kaikilla mitatuilla tekstipituuksilla. Kuvasta 3 voidaan nähdä, että ero Ranking- ja LLM-menetelmien suoriutumisen välillä on edellistä vertailua suurempi, joten myös muut erot eri menetelmien välillä ovat tilastollisesti merkittäviä.

Ranking- ja LLM-menetelmien suoriutumista 60 kielen joukossa vertailtiin laskemalla ensin p-arvot kullekin tekstipituudelle parittaisella t-testillä. Arvot vaihtelivat välillä 0,003–0,95. Yhdistetty p-arvo oli 0,36, joka ei ole riittävän pieni nollahypoteesin hylkäämiseksi. Voidaan siis todeta, että menetelmien suoriutumisessa ei ole ti-

lastollisesti merkittävää eroa. Vertailtaessa KNLM- ja Ranking-menetelmiä p-arvot vaihtelivat välillä $8,2\times10^{-6}$–$2,0\times10^{-2}$ ja yhdistetyksi p-arvoksi saatiin $1,4\times10^{-15}$, mikä oikeuttaa nollahypoteesin hylkäämiseen. KNLM-menetelmä suoriutuu siis tilastollisesti merkitsevästi Ranking-menetelmää paremmin. Kuvasta 4 voidaan nähdä, että Ranking-menetelmä suoriutuu hieman LLM-menetelmää paremmin, joten voidaan todeta, että myös ero KNLM- ja LLM-menetelmien välillä on tilastollisesti merkittävä.

4 Pohdintaa ja kehitysehdotuksia

Tässä tutkimuksessa saadut tulokset osoittavat, että kehittyneillä tasoitusmenetelmillä voidaan parantaa kielimallipohjaisen kielentunnistuksen tarkkuutta. Tutkimuksessa ei kuitenkaan otettu huomioon tunnistuksen vaatimaa laskennallista kustannusta. Yleisesti voidaan todeta, että vertailumenetelminä käytetyt LLM- ja Ranking-menetelmät ovat laskennallisesti hieman nopeampia ja niiden suorituskyky lähestyy KNLM-menetelmää tunnistettavan tekstilohkon pituuden kasvaessa. Saaduilla tuloksilla on siis todellista merkitystä vain jos halutaan tunnistaa kieli lyhyen syötteen perusteella.

Tutkimuksessa saadut erot Dunningin LLM-menetelmän ja KNLM-menetelmän välillä syntyvät siitä, ettei LLM-menetelmä ota lyhyempiä n-grammeja lainkaan huomioon. Kohdattaessa n-grammi, jota ei ollut opetusaineistossa, LLM-menetelmä turvautuu käyttämään tasoituksesta saatua arviota $1/|A|$, jossa A on kielen merkistön koko. Tämän vuoksi kaikki kohtaamattomat n-grammit saavat karkeasti yliarvioidut todennäköisyydet. Ongelma korostuu kielillä, joilla on suuri merkistö.

Taulukoista 2 ja 3 voidaan nähdä, että kaikki menetelmät toimivat suhteellisen huonosti tunnistettaessa englannin- ja espanjankielisiä näytteitä. Englannin tuloksia heikensi Etelä-Afrikassa puhutun vendan kielen materiaali, joka sisälsi myös englanninkielisen version ihmisoikeusjulistuksesta. Tämä huomattiin vasta kun kokeiden uusiminen tähän opinnäytteeseen oli liian myöhäistä. Englanti sekoittuu usein myös skotin kieleen, jota voidaan pitää englannin murteena. Espanja sekoittuu useimmin asturian ja galician kieliin, jotka ovat romaanisia, espanjassa puhuttuja vähemmistökieliä.

Google AJAX language APIn tulosten perusteella voidaan epäillä, että se on säädetty tunnistamaan yleisiä kieliä harvinaisempien kielten kustannuksella. Tämä selittää myös Googlen muita menetelmiä heikommat tulokset, vaikka taulukon 4 perusteella Google näyttäisi tunnistavan monia kieliä jopa muita menetelmiä paremmin.

Jatkossa olisi mielenkiintoista laajentaa vertailua koskemaan useampia menetelmiä. Erityisen mielenkiintoista olisi vertailla KNLM-menetelmän ja Teahanin ja Harperin (2001) käyttämien PPM-pakkausmenetelmien suoriutumista kielentunnistuksessa. Kummankin menetelmän voidaan tulkita mittaavan, kuinka monta bittiä merkkiä kohden tarvitaan tunnistettavan viestin koodaamiseen kullakin kielimallilla. Menetelmien erot ovat ainoastaan siinä, kuinka kielimallit rakennetaan.

N-grammimalleihin pohjautuvia kielentunnistimia olisi mahdollista parantaa ottamalla kutakin kielimallia rakennettaessa huomioon myös kaikki tunnetut kielet. Řehůřek ja Kolkus (2009) ovat pyrkineet arvioimaan eri sanojen merkitsevyyttä kussakin kielessä ottamalla huomioon kaikista kieliaineistoista koostuneen tausta-aineiston. Samankaltaista lähestymistapaa voitaisiin käyttää myös n-grammien tasolla. Vastaavaan tulokseen voitaisiin päästä myös laskemalla tausta-aineistosta erillinen taustakielimalli, jota käytettäisiin kielimallien antamien todennäköisyyksien normalisointiin. Parannusta voitaisiin saada myös rakentamalla normaalien kielimallien lisäksi takaperin käännettyihin aineistoihin perustuvat kielimallit, joita käytettäisiin hyväksi todennäköisyyksien arvioinnissa.

5 Yhteenveto

Tässä opinnäytetyössä tutkittiin kokeellisesti kielentunnistusmenetelmien toimintaa lyhyillä, 5–21 merkin mittaisilla tekstilohkoilla. Lisäksi pyrittiin motivoimaan puheentunnistuksessa käytetyillä menetelmillä tasoitettujen merkkipohjaisten n-grammimallien käyttöä kielentunnistukseen. Kokeellisen osuuden aineistona käytettiin YK:n ihmisoikeusjulistusta, mikä mahdollisti 298 kielen testijoukon.

Menetelmäosiossa pyrittiin kuvaamaan n-grammimallien tasoitusmenetelmien kehityskaari sekä selvittämään lukijalle nykyaikaisen Kneser–Ney-tasoituksen toimintaa käytettäessä merkkipohjaisia n-grammimalleja. Kielimallipohjaiset kielentunnistimet liitettiin lopulta informaatioteorian avulla ristientropiaa vertaileviin menetelmiin. Osiossa myös esiteltiin työssä käytetyt tilastolliset testit ja kokeellisessa osuudessa mukana olleet kielentunnistimet.

Kokeellisessa osuudessa testattiin kolmen kielentunnistusmenetelmän toimintaa 298 kielen joukossa. Menetelmiä olivat työssä esitelty Kneser–Ney-tasoitettuihin n-grammimalleihin perustuva KNLM-luokitin, Dunningin (1994) esittelemä yksinkertaisempiin n-grammimalleihin perustuva LLM-menetelmä sekä Cavnarin ja Trenklen (1994) kehittämä Ranking-menelmä. Lisäksi käytettiin karsittua 60 kielen testijoukkoa edellä mainittujen menetelmien vertaamiseksi suppeamman kielivalikoiman omaavaan Google AJAX Language APIin.

Kokeissa saadut tulokset osoittavat, että kehittyneillä tasoitusmenetelmillä voidaan parantaa kielimallipohjaisen kielentunnistuksen tarkkuutta. Työssä ehdotettu KNLM-menetelmä suoriutui kaikkia vertailumenetelmiä tilastollisesti merkittävästi paremmin sekä 60 kielen rajoitetulla että 298 kielen täydellä testiaineistolla.

Viitteet

Alpaydin, E. (2004). *Introduction to machine learning.* MIT Press, Cambridge, Massachusetts.

Botha, G. ja Barnard, E. (2007). Factors that affect the accuracy of text-based language identification. Kirjassa *18th Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*, ss. 7–12.

Cavnar, W. B. ja Trenkle, J. M. (1994). N-gram-based text categorization. Kirjassa *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, ss. 161–175.

Chen, S. F. ja Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–393.

Cleary, J. G., Ian, ja Witten, I. H. (1984). Data compression using adaptive coding and partial string matching. *IEEE Transactions on Communications*, 32:396–402.

Creutz, M. ja Lagus, K. (2007). Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, 4(1):1–34.

da Silva, J. F. ja Lopes, G. P. (2006). Identification of document language is not yet a completely solved problem. Kirjassa *International Conference on Computational Inteligence for Modelling Control and Automation (CIMCA 2006)*, ss. 212–219.

Damashek, M. (1995). Gauging similarity with n-grams: Language-independent categorization of text. *Science*, 267(5199):843–849.

Dunning, T. (1994). Statistical identification of language. Tekninen raportti MCCS-94-273, Computing Research Lab, New Mexico State University.

Fisher, R. A. (1948). Combining independent tests of significance (in response to question 14). *American Statistician*, 2(5):559–574.

Gold, E. M. (1967). Language identification in the limit. *Information and Control*, 10(5):447–474.

Goodman, J. (2008). A bit of progress in language modeling, extended version. Tekninen raportti MSR-TR-2001-72, Microsoft Research.

House, A. S. ja Neuburg, E. P. (1977). Toward automatic identification of the language of an utterance. I. Preliminary methodological considerations. *Journal of the Acoustical Society of America*, 62(3):708–713.

Jurafsky, D. ja Martin, J. H. (2008). *Speech and Language Processing.* Prentice Hall, 2nd edition.

Kneser, R. (1996). Statistical language modeling using a variable context length. Kirjassa *The Fourth International Conference on Spoken Language Processing*, volume 1, ss. 494–497.

Kneser, R. ja Ney, H. (1995). Improved backing-off for m-gram language modeling. Kirjassa *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, ss. 181–184.

Kruengkrai, C., Srichaivattana, P., Sornlertlamvanich, V., ja Isahara, H. (2005). Language identification based on string kernels. Kirjassa *Proceedings of the 5th International Symposium on Communications and Information Technologies (ISCIT-2005*, ss. 896–899.

Manning, C. D. ja Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press, 1 edition.

McNamee, P. (2005). Language identification: a solved problem suitable for undergraduate instruction. *Journal of Computing Sciences in Colleges*, 20(3):94–101.

Milton, J. S. ja Arnold, J. C. (2002). *Introduction to Probability and Statistics: Principles and Applications for Engineering and the Computing Sciences*. McGraw-Hill.

Ney, H., Essen, U., ja Kneser, R. (1994). On structuring probabilistic dependence in stochastic language modelling. *Computer Speech and Language*, 8(1):1–38.

Řehůřek, R. ja Kolkus, M. (2009). Language identification on the web: Extending the dictionary method. Kirjassa *Proceedings of CICLing 2009*, ss. 357–368.

Rosenfeld, R. (1996). A maximum entropy approach to adaptive statistical language modeling. *Computer, Speech and Language*, 10:187–228.

Rosenfeld, R. (2000). Two decades of statistical language modeling: Where do we go from here? Kirjassa *Proceedings of the IEEE*, numero 8, ss. 1270–1278.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656.

Sibun, P. ja Reynar, J. C. (1996). Language identification: Examining the issues. Kirjassa *Fifth Annual Symposium on Document Analysis and Information Retrieval (SDAIR'96)*, ss. 125–135.

Siivola, V., Hirsimäki, T., ja Virpioja, S. (2007). On growing and pruning Kneser-Ney smoothed n-gram models. *IEEE Transactions on Audio, Speech & Language Processing*, 15(5):1617–1624.

Souter, C., Churcher, G., Hayes, J., Hughes, J., ja Johnson, S. (1994). Natural language identification using corpus-based models. *Hermes Journal of Linguistics*, 13:183–203.

Teahan, W. J. (2000). Text classification and segmentation using minimum cross-entropy. Kirjassa *Computer-Assisted Information Retrieval (Recherche d'Information et ses Applications, RIAO)*, ss. 943–961.

Teahan, W. J. ja Harper, D. J. (2001). Combining PPM models using a text mining approach. Kirjassa *Data Compression Conference*, ss. 153–162.

Tuomainen, K. (2008). Kielen tunnistus tilastollisin menetelmin. Kandidaatintyö, Teknillinen Korkeakoulu, Informaatio- ja luonnontieteiden tiedekunta, Tietotekniikan tutkinto-ohjelma.

Virpioja, S. (2005). New methods for statistical natural language modeling. Diplomityö, Helsinki University of Technology, Department of Computer Science and Engineering, Laboratory of Computer and Information Science.

Zissman, M. A. (1996). Comparison of four approaches to automatic language identification of telephone speech. *IEEE Transaction on Speech and Audio Processing*, 4(1).