Aleksi Pekkala
aleksi.v.a.pekkala@student.jyu.fi

Sanaluokkien automaattisen tunnistamisen menetelmät

Tietotekniikka
TIEA217 Tietojenkäsittelyn alan tutkimusmenetelmät
30. lokakuuta 2013
Tentaattori: Hannakaisa Isomäki

Jyväskylän yliopisto
Tietotekniikan laitos

1 Johdanto

1.1 Mitä on sanaluokkien automaattinen tunnistaminen?

Sanaluokkien automaattisella tunnistamisella (*Part-of-speech tagging, POS tagging*) tarkoitetaan sanojen sanaluokkien tunnistamista tekstiyhteyden perusteella. Sanaluokittelun tavoitteena on rakentaa malli, joka ottaa syötteenään lauseen, kuten

$$a\ black\ cat\ jumped\ on\ the\ table$$

ja jonka tuloksena on sarja sanaluokkia, jotka vastaavat annettua lausetta, tässä tapauksessa esim.

$$Det\ Adj\ Noun\ Verb\ Prep\ Det\ Noun$$

Sanaluokkien tunnistaminen on luonteeltaan sarjanluokitteluongelma (*sequence labeling problem*), joka puolestaan kuuluu hahmontunnistuksen (*pattern recognition*) ongelmatyyppeihin.

1.2 Mihin sanaluokkien tunnistamista käytetään?

Sanaluokkien tunnistaminen on tärkeä ja käytännöllinen ongelma, joka kohdataan useimmissa luonnollisten kielten käsittelyyn liityvissä tehtävissä, joihin kuuluvat mm. puheentunnistus, konekääntäminen, semanttinen haku sekä automaattinen vastaaminen. Tällaisissa tehtävissä sanaluokkien tunnistaja toimii jonkin laajemman prosessointiketjun alkupäässä, koko prosessille välttämättömänä esikäsittelijänä, joka mahdollistaa syötteen jatkokäsittelyn korkeammalla tasolla. Jatkokäsittelyn tyypillisin seuraava vaihe on tekstin jäsentäminen eli lauserakenteiden tunnistaminen, jonka onnistumisen kannalta on oleellista, että sanaluokkien tunnistaminen on suoritettu mahdollisimman virheettömästi.

1.3 Sanaluokkien tunnistajan vaatimukset

Jotta tunnistajaa voidaan käyttää laajan kielenprosessointijärjestelmän komponenttina, tulee sen toteuttaa seuraavat ominaisuudet (Cutting ym., 1992):

Kestävyys Tunnistajan tulee kyetä selviytymään kaikista tekstisyötteen mahdollisista poikkeamista, kuten otsikoista, taulukoista sekä tuntemattomista sanoista.

Tehokkuus Voidakseen käsitellä laajoja tekstiaineistoja tunnistajan tulee toimia lineaarisessa ajassa.

Tarkkuus Tunnistajan tulee kyetä ehdottamaan sanaluokka jokaista annettua sanaa kohden. Myös tunnistajan mahdollisen opettamisen tulisi onnistua mahdollisimman nopeasti.

Viritettävyys Tunnistajan tulee osata hyödyntää erilaisia kielitieteellisiä huomioita siten, että tunnistajan tekemiä virheitä voidaan paikata määrittämällä sopivia vihjeitä.

Uudelleenkäytettävyys Tunnistajan tulee rakentua siten, että sen kohdistaminen uudelle kielelle, aineistolle tai sanaluokkasetille on mahdollisimman vaivatonta.

1.4 Miksi sanaluokkien tunnistaminen on ongelmallista?

Sanaluokkien tunnistaminen voi intuitiivisesti tuntua helpolta, mutta tehtävän automaatiota hankaloittavat kaksi oleellista ongelmaa: sanojen monitulkintaisuus sekä tuntemattomat sanat. Esimerkiksi lauseet

$$Time\ flies\ like\ an\ arrow$$

$$Fruit\ flies\ like\ a\ banana$$

voidaan tulkita lukuisin eri tavoin, joista mikään ei ole välttämättä muita ilmeisempi. Lisäksi, vaikka tosielämässä tulkittavat lauseet ovat harvoin yhtä ongelmallisia kuin edellämainitut lingvistiset esimerkkilauseet, on monitulkintaisuus hyvin yleistä: arviolta 40% englanninkielisen proosan sanastosta omaa useamman kuin yhden merkityksen (DeRose, 1988). Monitulkintaisuuden ollessa ongelman keskiössä olisikin ehkä luontevampaa puhua sanaluokkien tunnistamisen sijaan yksikäsitteistämisestä (*disambiguation*). On myös mielenkiintoista huomata, että edes ammattilaisten käsin luokittelemat aineistot eivät saavuta täydellistä tunnistamistarkkuutta: ihmisten sanaluokkien tunnistamistarkkuuden on arvioitu olevan noin 97%, mikä vastaa nykyisten automaattisten tunnistajien huipputuloksia (Manning, 2011).

Toinen merkittävä ongelma on tuntemattomien, eli harjoitusaineistosta puuttuvien sanojen käsitteleminen. Tuntemattomia sanoja kohdataan usein, ja koska niitä koskevaa tilastollista informaatiota tai sääntöjä ei tunneta, joudutaan turvautumaan jonkinlaiseen poikkeuskäsittelyyn. Useat tunnistajat hyödyntävät tuntemattomien sanojen kohdalla kieliopillisia ominaisuuksia: yksinkertainen menetelmä on määrätä sanalle se sanaluokka, joihin tuntemattomien sanojen on havaittu todennäköisimmin kuuluvan, eli yleensä substantiivi. Parempia tuloksia on saavutettu määrittämällä tuntemattoman sanan sanaluokka sen päätteen perusteella; esim. englannin kielen *able*-päätteiset sanat ovat hyvin todennäköisesti adjektiiveja (Samuelsson, 1993). Menetelmä ei kuitenkaan sovellu kaikille kielille: esim. Tseng ym. (2005) osoittavat, että kiinan kielessä esiintyy huomattavan suuri määrä yleisiä affikseja, poiketen englannin ja saksan kielistä, joissa affiksit ovat vahva indikaattori sanan sanaluokasta.

Ongelmallista on myös tunnistamisessa käytettävien sanaluokkien (*POS tagset*) valinta. Yleisiä englannin kielen sanaluokkasettejä ovat Brownin aineiston 87 sanaluokkaa (Francis,

1964), tai uudemman Penn Treebank-aineiston 48 sanaluokkaa (Marcus ym., 1993). Myös yleisiä, kielestä riippumattomia sanaluokkasettejä on kehitetty, joskin tällöin joudutaan väistämättä tinkimään tunnistamistarkkuudesta (Petrov ym., 2011).

1.5 Automaattisten tunnistajien suorituskyky

Kuten mainittua, nykyisten automaattisten sanaluokkien tunnistajien tunnistamistarkkuus on hieman yli 97% (Toutanova ym., 2003, Shen ym., 2007, Spoustova ym., 2009, Søgaard, 2010). Manning (2011) kuitenkin osoittaa, että kyseistä tulosta ei ole syytä tulkita liian optimistisesti: esimerkiksi lukuisat välikemerkit ja muut yksikäsitteiset elementit vääristävät evaluaatiotuloksia. Artikkelissa huomautetaankin, että realistisempaa olisi tarkastella automaattisten tunnistajien kykyä tunnistaa kokonaiset lauseet oikein, sillä pienikin virhe lauseessa voi vahingoittaa tunnistajan hyödyllisyyttä myöhempien prosessointivaiheiden kannalta; tällä saralla tunnistajat saavuttavat noin 55-57% tarkkuuden, mikä on huomattavasti vaatimattomampi tulos.

2 Kirjallisuuskartoitus

2.1 Aineiston hakuprosessi

Tavoitteenani oli kuvailla viittä erilaista automaattista sanaluokkien tunnistamismenetelmää. Asetin menetelmille seuraavat vaatimukset:

Monipuolisuus Menetelmien tulee poiketa toisistaan merkittävästi.

Suorituskyky Jotta menetelmien vertailu on kannattavaa, tulee niiden olla keskenään kilpailukykyisiä. Tämä todetaan vertaamalla samalla aineistolla mitattuja tunnistustarkkuustuloksia.

Yleistettävyys Kunkin menetelmän tulee olla yleistettävissä eri kielille, aineistoille ja sanaluokkaseteille; menetelmä ei saa olla vain johonkin tiettyyn erityistapaukseen soveltuva. Menetelmän tulee olla toteutettu käytännössä.

Yleisyys Menetelmän tulee olla suhteellisen tunnettu ja laajalta tutkittu. Todetaan tarkastelemalla kirjallisuuden ja viittausten määrää.

Menetelmiä valitessa erinomaiseksi apuvälineeksi osoittautui ACL:n listaus nykyisistä ”state-of-the-art”-tunnistusmenetelmistä evaluaatiotuloksineen (”POS Tagging State of the Art”, 2013).

Löydettyäni tavoitteidenmukaiset tunnistamismenetelmät, etsin kutakin menetelmää kohti ainakin yhden artikkelin, joka sisälsi mahdollisimman tarkan kuvauksen kyseisestä menetelmästä. Lisäksi artikkelin vaatimuksena oli se, että se kuvaa tunnistamismenetelmänsä käytän-

nöllisesti sekä toteuttamiskelpoisena. Tämän seurauksena valittu artikkeli ei välttämättä ollut juuri se julkaisu, jossa menetelmä esiteltiin ensimmäistä kertaa, vaan osa artikkeleista kuvailee vanhan menetelmän pohjalta toteutettua modernimpaa tunnistajaa.

Artikkeleiden hakemiseen käytin Google- ja Google Scholar-hakukoneita, hakuehtona aina kunkin menetelmän nimi. Tarkastelin tarkemmin 21 eri artikkelia, joista valitsin lopulta 5 tärkeintä, lähinnä artikkelien välisten viittausten perusteella. Suurin osa artikkeleista löytyi ACM:n verkkokirjastosta.

2.2 Kuvaus käytetyistä artikkeleista

2.2.1 TnT - A Statistical Part-of-Speech Tagger, Brants (2000)

Brants (2000) esittää artikkelissaan tilastollisen Trigrams'n'Tags-tunnistajan (TnT), joka perustuu Markovin piilomalleihin. Markovin piilomalleja on käytetty aikaisemmissakin tunnistimissa (mm. DeRose, 1988, Cutting ym., 1992, Merialdo, 1994), joiden ideoita jalostamalla TnT-tunnistin saavuttaa huomattavan korkean tunnistamistarkkuuden. Tunnistajan toiminnan esitetään artikkelissa perustuvan siihen, että lasketaan

$$\arg\max_{t_1,\ldots,t_T} \left[\prod_{i=1}^{T} P(t_i|t_{i-1}, t_{i-2}) P(w_i|t_i) \right] P(t_{T+1}|t_T)$$

missä $w_1, \ldots, w_T$ ovat sanoja ja $t_1, \ldots, t_T$ sanaluokkia. Todennäköisin sanaluokkasarja $t_1, \ldots, t_T$ ratkaistaan dynaamisella Viterbi-algoritmilla (Rabiner, 1989). Kohdassa

$$P(t_i|t_{i-1}, t_{i-2})$$

lasketaan todennäköisyys sanaluokalle t_i, kun kaksi edeltävää sanaluokkaa ovat t_{i-1} ja t_{i-2}. Tästä huomataan, että kyseessä on toisen asteen Markovin piilomalli - lisäksi tunnistimen nimi viittaa kyseiseen trigrammiin. Kohdassa

$$P(w_i|t_i)$$

lasketaan todennäköisyys sille, että sana w_i esiintyy sanaluokan t_i yhteydessä, ja lopun

$$P(t_{T+1}|t_T)$$

merkitsee todennäköisyyttä sille, että päätemerkkisanaluokka t_{T+1} seuraa sarjan viimeistä sanaa.

Artikkelissa kuvataan myös silotusmenetelmä, jolla voidaan lievittää liian pienestä aineistosta johtuvia vääristymiä. Lisäksi tunnistajaan kuuluu mm. tuntemattomien sanojen käsittely

suffiksianalyysin (Samuelsson, 1993) avulla, sekä isojen alkukirjainten hyödyntäminen sanaluokkien tunnistamisessa. Lopuksi artikkelissa ilmoitetaan vielä Penn Treebank-aineistolla saavutettu tunnistamistarkkuus 96,7%.

2.2.2 A Simple Rule-Based Part of Speech Tagger, Brill (1992)

Brill (1995) huomauttaa, että edeltävän kaltaiset tilastolliset tunnistajat saavuttavat korkean tunnistamistarkkuuden kiinnittämättä varsinaisesti huomiota aineiston taustalla olevaan kieliopilliseen rakenteeseen. Hän laskeekin tilastollisten menetelmien puutteeksi sen, että ne poimivat aineistosta lingvistisen informaation sijaan vain suuren määrän vaikeselkoisia tilastoja. Brill (1992) esittää artikkelissaan vaihtoehtoisen lähestymistavan, joka pohjautuu aineistosta havaittuihin kieliopillisiin sääntöihin. Aikaisemmista sääntöpohjaisista tunnistamismenetelmistä (mm. Garside, 1987) poiketen sääntöjä ei syötetä manuaalisesti, vaan tunnistin poimii ne automaattisesti annetusta aineistosta. Sääntöjen tunnistaminen perustuu siihen, että käytössä olevaa, oikeilla sanaluokilla merkittyä aineistoa voidaan käyttää valvottuun oppimiseen.

Säännöt selvitetään siten, että aineiston sanat merkitään ensin yksinkertaisesti sillä sanaluokalla, joka aineistossa esiintyy useimmiten annetun sanan kohdalla. Tämän jälkeen itse merkattua aineistoa verrataan oikein merkittyyn aineistoon, ja muodostetaan sääntöjä virheellisesti tunnistettujen sanaluokkien kohdalla. Muodostetuista säännöistä valitaan paras (arvon *oikeat korjaukset - virheelliset korjaukset* mukaan), sovelletaan sitä aineistoon, ja toistetaan halutun sääntömäärän mukaan. Säännöt voivat olla esimerkiksi muotoa

$$korvataan\ adjektiivi\ substantiivilla,\ jos\ edellinen\ sana\ on\ banana, \text{ tai}$$

$$korvataan\ pronomini\ konjunktiolla,\ jos\ seuraavan\ sanan\ sanaluokka\ on\ verbi$$

Artikkelin mukaan sääntöpohjaisella lähestymistavalla on monia etuja tilastollisiin menetelmiin verrattuna: tallennettavan informaation pieni koko ja luettavuus, parannuskohteiden löytämisen ja toteuttamisen helppous sekä parempi siirrettävyys aineistolta tai sanaluokkasetiltä toiselle.

2.2.3 A Maximum Entropy Model for Part-of-Speech Tagging, Ratnaparkhi (1996)

Ratnaparkhi (1997) huomauttaa raportissaan, että monet luonnollisten kielten käsittelyyn liittyvät ongelmat voidaan esittää tilastollisena luokitteluongelmana, jossa tehtävänä on arvioida ”luokan” a esiintymistä ”kontekstin” b yhteydessä, eli $p(a, b)$, missä kontekstilla tarkoitetaan tässä yhteydessä tiettyä sanaa ja sitä ympäröiviä sanoja sanaluokkineen. Suurimmatkaan aineistot eivät kuitenkaan sisällä tarpeeksi informaatiota määrittääkseen $p(a, b)$ jokaiselle parille (a, b), koska sanat b ovat usein harvinaisia; tällöin ongelmana on löytää menetelmä, jolla voidaan käytössä olevaa harvaa informaatiota hyödyntämällä arvioida luotettava todennäköisyysmalli $p(a, b)$.

Raportti ehdottaa ratkaisuksi suurimman entropian periaatetta (Jaynes, 1957), missä entropialla tarkoitetaan sitä, kuinka selvästi havaitut muuttujan arvot keskittyvät yhteen tai vain muutamaan luokkaan (entropia on suurimmillaan silloin, kun vaihtelu on suurinta). Kyseisen periaatteen mukaan oikea jakauma $p(a, b)$ on se, joka saavuttaa suurimman entropian kun huomioidaan ne ominaisuudet, joita käytettävistä muuttujista tiedetään.

Ratnaparkhi (1996) kuvaa aikaisemmassa artikkelissaan kuinka suurimman entropian periaatetta voidaan hyödyntää sanaluokkien tunnistamisongelman ratkaisussa, saavuttaen tunnistamistarkkuuden 96,6%. Artikkelin mukaan suurimman entropian periaatetta käyttävä malli soveltuu sanaluokkien tunnistamiseen erityisen hyvin, koska se pystyy hyödyntämään erilaista kontekstuaalista informaatiota tehokkaasti, eikä se ota kantaa harjoitusaineiston sana/sanaluokkaparien mahdollisesti epätasaiseen jakaumaan.

2.2.4 *Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network, Toutanova ym. (2003)*

Toutanova ym. (2003) arvostelevat aikaisempia tilastollisia tunnistimia (esim. Brants, 2000) siitä, että ne hyödyntävät lauseita käsitellessään vain yksisuuntaista informaatiota: esimerkiksi ensimmäisen asteen Markovin piilomallissa, jonka suunta on vasemmalta oikealle, nykyinen sanaluokka t_i arvioidaan edeltävän sanaluokan t_{i-1} sekä nykyisen sanan w_i perusteella, eli $P(t_i|t_{i-1}, w_i)$.

Artikkelissa yksisuuntaisuuden haittavaikutusta havainnollistetaan seuraavalla esimerkillä: sanalla *to* on vain yksi mahdollinen sanaluokka (TO). Sanaluokkaa TO edeltää usein substantiivi, mutta harvoin modaaliverbi. Lauseessa *will to fight* kyseisen säännön mukaan sanan *will* tulee olla substantiivi, ei modaaliverbi. Käytännössä sana *will* tunnistetaan kuitenkin modaaliverbiksi, koska käytettävä tilastollinen malli huomio *will*-sanan kohdalla vain edeltävän sanaluokan ja nykyisen sanan, ei seuraavan sanan sanaluokkaa.

Artikkelissa ongelmaan esitetään vaihtoehtoinen, syklisiin riippuvuusverkkoihin (Heckerman ym., 2000) perustuva menetelmä, joka ei Markovin ketjun tavoin rajoita solmujen riippuvuussuhteita yksisuuntaisiksi. Lisäksi artikkelissa kuvaillaan tuntemattomien sanojen käsittely sekä silotusmenetelmä, joita käyttämällä tunnistajan tarkkuudeksi lasketaan 97.24%.

2.2.5 *Fast and Accurate Part-of-Speech Tagging: The SVM Approach Revisited, Giménez & Màrquez (2003)*

Gimenéz & Màrquez (2003) esittelevät artikkelissaan tukivektorikoneisiin (*Support Vector Machine*) perustuvan sanaluokkatunnistajan. Tukivektorikone on koneoppimisalgoritmi, joka soveltuu binääriseen luokitteluun. Tukivektorikone opetetaan harjoitusesimerkeillä $\{(x_1, y_1), \ldots, (x_N, y_N)\}$, missä kukin x_i on vektori $\in \mathbb{R}^N$, ja $y_i \in \{-1, +1\}$ kuvaa vektorin luokkaa. Harjoitusaineiston perusteella tukivektorikone laskee tason, joka erottaa luokat toisistaan mahdollisimman suurella

marginaalilla.

Koska sanaluokan tunnistamisessa ei ole kyse binäärisestä luokittelusta, ongelma tulee binärisoida. Artikkelissa binärisöinti on toteutettu harjoittamalla jokaista sanaluokka kohti oma tukivektorikone, joka jakaa aineiston kyseiseen sanaluokkaan ja kaikkiin muihin sanaluokkiin.

3 Tutkimusaihe/tutkimuskysymys

Sanaluokkien tunnistajia käytetään monissa erilaisissa luonnollisiin kieliin liittyvissä sovelluksissa. Lisäksi tunnistettavien aineistojen välillä on valtavasti poikkeamia, esim. kielten ja tekstityyppien osalta. Lisäksi havaitaan, että tarkasteltujen artikkeleiden raportoimat tunnistustarkkuudet liikkuvat kaikki samoissa lukemissa. Kun ilmiselvin valintakriteeri on näin poissuljettu, on tehokkaimman menetelmän valinta vaikeampaa. Tutkittavaa olisikin siinä, **miten eri menetelmät eroavat toisistaan, ja soveltuvatko jotkin menetelmät toisia paremmin tietyille toimintaympäristöille**. Täsmentäviä kysymyksiä ovat ainakin

- Mitkä ovat yleisimmät sanaluokkien tunnistamismenetelmät?
- Millaisiin ryhmiin tunnistamismenetelmät voidaan jaotella?
- Millaisia eroja sanaluokkatunnistajien toimintaympäristöillä on?
- Miten eri tunnistamismenetelmät käyttäytyvät erilaisten aineistojen yhteydessä?
- Kuinka eri menetelmät poikkeavat harjoitusaineiston käytössä?
- Onko eri menetelmillä saavutettujen virheellisten tunnistustulosten välillä poikkeamia?

4 Tutkimusstrategia/metodi ja sen valinta

Tutkimuksen tarkoituksena on kartoittaa tutkittavan aihealueen nykytilaa kokoavasti ja katsauksenomaisesti. Tutkimus toteutetaan analysoituna kirjallisuuskatsauksena. Kirjallisuuskatsaus sopii metodina tutkimuksen luonteeseen; lisäksi tutkimusaiheesta löytyy laajalti laadukasta aineistoa.

5 Aineiston keruun suunnittelu ml. eettiset näkökohdat

Aineisto kerätään tehdyn kirjallisuuskatsauksen pohjalta. Kirjallisuuskatsauksen viidestä tunnistamismenetelmästä valitaan tärkeimmät, joihin syvennytään tarkemmin. Aineiston keruussa

on huomioitava se, että tutkimusaihe on osa laajempaa aihe-aluetta: on selvitettävä kunkin menetelmän alkuperä, ei pelkästään menetelmän sovellukset tutkimusaiheen piirissä.

6 Tietojen keruu

Alan kirjallisuuden (mm. Cutting ym., 1992) perusteella voidaan hahmotellaan sanaluokkatunnistajien tärkeimmät ominaisuudet. Näitä ominaisuuksia hyödyntäen tarkastellaan kutakin eri menetelmää systemaattisesti. Lisäksi kerätään tietoja siitä, miten eri menetelmiä on hyödynnetty erilaisissa erikoistapauksissa, esimerkiksi erilaisten kielten tunnistamisessa - tässä voidaan hyödyntää erityisesti *forward search*-menetelmää, eri tarkastella niitä artikkeleita, joissa on viitattu alkuperäisesti menetelmän esittäneeseen artikkeliin.

7 Tietojen analysointi

Kerättyjen tietojen perusteella arvioidaan, kuinka eri menetelmät, suhteessa toisiin menetelmiin, käyttäytyvät eri olosuhteissa. Analyysia varten tulee saavuttaa selkeä käsitys menetelmien toiminnasta, jotta vertailu on oikeudenmukaista ja johtopäätökset perusteltuja. Tarkemmin analysoidaan ainakin kunkin menetelmän suhdetta aineiston tekstityyppiin, aineiston kieleen, harjoitusaineiston määrään ja laatuun sekä tunnistamisvirheisiin.

8 Tulosten julkaiseminen

Tulokset julkaistaan kirjallisuuskatsauksen muodossa.

9 Johtopäätökset

Johtopäätöksinä voidaan esittää ainakin hypoteeseja siitä, miksi jokin menetelmä voisi olla toista parempi tietyssä käyttötapauksessa. Tutkimukseen ei kuitenkaan kuulu empiiristä kokeilua (menetelmien toteuttamista ja testaamista oikeilla aineistoilla), minkä vuoksi johtopäätöksiä ei voida varsinaisesti validoida; tehtävät johtopäätökset pohjautuvat siis aikaisempaan kirjallisuuteen.

Kirjallisuutta

Brants, T. 2000. *TnT - A statistical part-of-speech tagger*. Proceedings of the 6th Applied NLP Conference (ANLP).

Brill, E. 1992. *A simple rule-based part of speech tagger*. Proceedings of the 3rd conference on Applied Computational Linguistics, ACL, Trento, Italy.

Brill, E. 1995. *Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging*. Computational Linguistics, 21(4), s. 543-565.

Church, K. 1988. *A stochastic parts program and noun phrase parser for unrestricted text*. Proceedings of the 2nd conference on Applied Natural Language Processing, s. 136-143.

Cutting, D., Kupiec, J., Pedersen, J. & Sibun, P. 1992. *A Practical Part-of-Speech Tagger*. Proceedings of the 3rd conference on Applied Natural Language Processing, s. 133–140.

DeRose, S. J. 1988. *Grammatical category disambiguation by statistical optimization*. Computational Linguistics, 14(1), s. 31–39.

Francis, W. N. 1964. *A standard sample of present-day English for the use with digital computers*. Report for the U.S Office of Education on Cooperative Research Project No. E-007. Brown University, Providence RI.

Garside, R. 1987. *The CLAWS word-tagging system*. Teoksessa R. Garside, G. Leech & G. Sampson (toim.) The Computational Analysis of English: A Corpus-based Approach. London: Logman, s. 30-41.

Gimenéz J. & Màrquez, L. 2003. *Fast and accurate part-of-speech tagging: The SVM approach revisited*. Proceedings of RANLP 2003, s. 158-165.

Heckerman, D., Chickering, D. M., Meek, C., Rounthwaite, R. & Kadie, C. M. 2000. *Dependency networks for inference, collaborative filtering and data visualization*. Journal of Machine Learning Research, 1(1), s. 49-75.

Jaynes, E. T. 1957. *Information Theory and Statistical Mechanics*. Physical Review, 106, s. 620-630.

Manning, C. D. 2011. *Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?* Proceedings of the 12th international conference on Computational linguistics and intelligent text processing, 1, s. 171–189.

Marcus, M. P., Marcinkiewicz, M. & Santorini, B. 1993. *Building a large annotated corpus of English: the Penn treebank*. Computational Linguistics, 19(2), s. 313-330.

Merialdo, B. 1994. *Tagging English text with a probabilistic model*. Computational Linguistics, 20(2), s. 155-171.

Petrov, S., Das, D. & McDonald, R. 2011. *A universal part-of-speech tagset*. ArXiv:1104.2086.

POS Tagging State of the Art. 2013. The Wiki of the Association for Computational Linguistics. Haettu 28.10.2013, osoitteesta `aclweb.org/aclwiki/index.php?title=POS_Tagging_(State_of_the_art)`

Rabiner, L. R. 1989. *A tutorial on Hidden Markov Models and selected applications in speech recognition*. Proceedings of the IEEE, 77(2), s. 257-285.

Ratnaparkhi, A. 1996. *A maximum entropy model for part-of-speech tagging*. Proceedings of the Empirical Methods in Natural Language Processing Conference (EMNLP), University of Pennsylvania.

Ratnaparkhi, A. 1997. *A simple introduction to maximum entropy models for natural language processing*. Technical Report 97-08, Institute for Research in Cognitive Science, University of Pennsylvania, 1997.

Samuelsson, C. 1993. *Morphological tagging based entirely on Bayesian inference*. Proceedings of the 9th Nordic Conference on Computational Linguistics NODALIDA-93.

Shen, L., Satta, G. & Joshi, A. 2007. *Guided learning for bidirectional sequence classification*. In: ACL 2007. (2007)

Spoustova, D.j., Hajic, J., Raab, J. & Spousta, M. 2009. *Semi-supervised training for the averaged perceptron POS tagger*. Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), s. 763-711.

Søgaard, A. 2010. *Simple semi-supervised training of part-of-speech taggers*. Proceedings of the ACL 2010 Conference Short Papers, s. 205-208.

Toutanova, K., Klein, D., Manning, C.D. & Singer, Y. 2003. *Feature-rich part-of-speech tagging with a cyclic dependency network*. In: NAACL 3. (2003), s. 252-259

Tseng, H., Jurafsky, D. & Manning, C. 2005. *Morphological features help POS tagging of unknown words across language varities*. Proceedings of the 4th SIGHAN bakeoff.