Aleksi Pekkala

Sanaluokkien automaattisen tunnistamisen menetelmät

Tietotekniikan kandidaatintutkielma

27. marraskuuta 2013

Jyväskylän yliopisto

Tietotekniikan laitos

Tekijä: Aleksi Pekkala

Yhteystiedot: aleksi.v.a.pekkala@student.jyu.fi

Työn nimi: Sanaluokkien automaattisen tunnistamisen menetelmät

Title in English: Methods for automated part-of-speech tagging

Työ: Kandidaatintutkielma

Sivumäärä: 23+0

Tiivistelmä: Tiivistelmä on tyypillisesti 5-10 riviä pitkä esitys työn pääkohdista (tausta, tavoite, tulokset, johtopäätökset).

Avainsanat: kieliteknologia, luonnollisten kielten käsittely, sanaluokkien tunnistaminen, koneoppiminen

Abstract: Englanninkielinen versio tiivistelmästä.

Keywords: computational linguistics, natural language processing, part-of-speech tagging, machine learning

Sisältö

1 Johdanto

Sanaluokkien automaattinen tunnistaminen (engl. *part-of-speech tagging*) tarkoittaa sanojen sanaluokkien tunnistamista tekstiyhteyden perusteella. Tunnistamisprosessissa tarkastellaan tekstiaineistoa, kuten

a black cat jumped on the table

jonka perusteella pyritään päättelemään se sanaluokkien sarja, joka todennäköisimmin vastaa kyseistä aineistoa; tässä tapauksessa tuloksena voisi olla esimerkiksi

Det Adj Noun Verb Prep Det Noun

Sanaluokkien tunnistaminen on laajuudeltaan rajallinen ongelma: sen tarkoituksena ei ole jäsentää kokonaisia lauserakenteita tai tulkita lauseiden merkitystä — tarkastelun alla ovat vain yksittäisten sanojen syntaktiset kategoriat. Sanaluokkien tunnistaminen on kuitenkin välttämätön ensimmäinen askel useimmissa luonnollisten kielten käsittelyprosesseissa, ja siten yksi aihealueen keskeisimmistä osaongelmista.

Rajallisen laajuutensa myötä sanaluokkien tunnistaminen on paljon helpommin lähestyttävä ongelma kuin kielen täydellinen ymmärtäminen, ja sen ratkaisemiseksi onkin kehitetty useita kohtuullisen luotettavia menetelmiä. Täysin ratkaistusta ongelmasta ei kuitenkaan voida puhua, sillä yksikään tunnettu menetelmä ei vielä saavuta täydellistä tunnistustarkkuutta.

Sanaluokkien tunnistajia käytetään monissa erilaisissa luonnollisiin kieliin liittyvissä sovelluksissa, ja tunnistajalle asetetut vaatimukset vaihtelevat sovelluksittain. Myös tunnistettavien aineistojen välillä on valtavasti poikkeamia, esim. kielten sekä tekstilajien osalta. Lisäksi havaitaan, että nykyisten tunnistusmenetelmien saavuttamat tunnistustarkkuudet liikkuvat kaikki suunnilleen samoissa lukemissa. Kun ilmiselvin valintakriteeri on näin poissuljettu, on tehokkaimman menetelmän valin-

ta vaikeampaa. Tunnistusmenetelmien toimintaperiaatteiden vaihdellessa merkittävästi on kuitenkin väistämätöntä, että jotkin menetelmät soveltuvat toisia paremmin tiettyihin tunnistustehtäviin. Tässä tutkielmassa pyritäänkin selventämään sitä, ~~kuinka eri tunnistusmenetelmät käyttäytyvät suhteessa toisiinsa erilaisissa toimintaympäristöis~~ millaisia eri ratkaisuja sanaluokkien tunnistusongelmaan on olemassa ja mitkä ovat niiden ~~oleellisimmat vahvuudet sek10001000~~t10001000~~heikkoudet~~rkeimmät erot. Menetelmien suhteelliset ominaisuudet johdetaan tarkastelemalla lähemmin kunkin menetelmän toimintaa ~~,~~ sekä menetelmään liittyvää tutkimuskirjallisuutta.

Tutkielma rakentuu seuraavasti: toisessa luvussa annetaan lyhyt johdanto sanaluokkien tunnistamiseen ja sen haasteisiin. Luvuissa 3-5 tarkastellaan kolmea erilaista tunnistusmenetelmää~~,~~: transformaatiosääntöjä, Markovin piilomalleja sekä log-lineaarisia malleja. Luvuissa esitellään ~~niiden~~ menetelmien keskeiset ominaisuudet, ~~ja~~ sekä pyritään hahmottamaan ~~kunkin menetelm10001000~~niiden suhteelliset vahvuudet ja heikkoudet. Lopuksi viel10001000 ~~n suhteelliset vahvuudet . TODO mainitaan~~ kootaan yhteen menetelmistä kerätyt huomiot ja esitellään johtopäätökset.

TODO olisiko syytä esitellä lyhyesti valitut menetelmät ~~ja valintaperusteet (~~tässä, ja/tai mainita jotain valintaperusteista tai esitysjärjestyksestä ~~:~~ (kaksi ensimmäistä ovat tavallaan toistensa vastakohtia, ja kolmas menetelmä yhdistää piirteitä kummastakin aikaisemmasta menetelmästä. Samalla menetelmät ovat järjestetty yksinkertaisimmasta monimutkaisimpaan.).

2 Sanaluokkien automaattinen tunnistaminen

2.1 ~~Mihin sanaluokkien tunnistamista käytetään?~~

Sanaluokkien tunnistaminen on tärkeä ja käytännöllinen ongelma, joka kohdataan ~~useimmissa~~ lähes kaikissa luonnollisten kielten käsittelyyn liittyvissä tehtävissä. Tällaisia tehtäviä ovat mm. puheentunnistus, konekääntäminen sekä semanttinen haku ja analyysi. Kyseisissä 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. Oikeiden lauserakenteiden tunnistamisen kannalta on oleellista, että lauseiden sanaluokat on tunnistettu mahdollisimman virheettömästi: yksikin virheellinen sanaluokka voi tehdä oikean lauserakenteen tunnistamisesta mahdotonta, ja siten vääristää lauseen tulkittua merkitystä.

2.1 Sanaluokkien tunnistamisen lyhyt historia

TODO

2.2 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 TODO back door, back of my foot -esimerkki, josta ilmenee sanaluokan monitulk.

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) .~~

voidaan luokitella useampaan kuin yhteen sanaluokkaan (DeRose, 1988) .

Jatkokäsittelyn kannalta automaattisen tunnistajan oleellisin tehtävä onkin ~~sen sopivimman merkityksen valinta, eli ns. morfologinen~~ valita kaikista mahdollisista sanaluokista se, joka tuottaa luontevimman tulkinnan. Tällaisen yksikäsitteistä~~minen. Yksikäsitteistä~~misen mahdollistavat luonnollisten kielten sisäänrakennetut rajoitteet, ~~ja erityisesti kaksi oleellista vihjetyyppi10001000~~jotka voidaan jakaa lokaaleihin sek10001000 ~~: lokaalit vihjeet~~ kontekstuaalisiin vihjeisiin: lokaaleista vihjeistä ilmeisin on itse sana ("sana *can* on on todennäköisemmin modaaliverbi kuin substantiivi"), mutta päätelmiä voidaan tehdä myös esimerkiksi sanan prefiksin, suffiksin sekä ~~kontekstuaaliset vihjeet (~~"alkukirjaimen koon perusteella. Kontekstuaalisia vihjeitä ovat kaikki lauseen muut sanat sanaluokkineen: esimerkiksi sana *fly* on todennäköisimmin substantiivi, jos edeltävä sana on artikkeli~~")~~.

On tärkeää huomata, ettei sanaluokkien tunnistaminen itsessään ole ratkaisu kieliopilliseen monitulkintaisuuteen: monitulkintaisuudella on useita tasoja, joista osaa käsitellään vielä prosessointiketjun myöhemmissä vaiheissa. Esimerkiksi syntaktinen, tai rakenteellinen monitulkintaisuus on ongelma, joka on huomattavasti helpompi ratkaista lauseiden jäsennyksen yhteydessä. Sanaluokkien tunnistamista ei myöskään tule sekoittaa semanttiseen yksikäsitteistämiseen, eli sanan merkityksen selvittämiseen: esimerkiksi sana *mouse* on semanttisesti monitulkintainen, vaikka sen sanaluokka tunnettaisiinkin. Sanaluokkien tunnistusprosessin rooli on pikemminkin rajata mahdollisten tulkintojen määrää prosessointiketjun alkupäässä, jotta myöhemmissä vaiheissa vältytään turhalta työltä.

Toinen merkittävä ongelma on tuntemattomien, eli harjoitusaineistosta puuttuvien sanojen käsitteleminen ~~. Tuntemattomia sanoja~~ (TODO: onko tämä ymmärrettävä selitys? harjoitusaineistosta tai ohjatusta oppimisesta ei ole vielä mainittu mitään). Englannin kielessä yleisiä tuntemattomia sanoja ovat erisnimet sekä puhekieliset,

vieraskieliset ja muut harvinaiset ilmaisut. Tällaisia 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 (engl. *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) .~~

2.3 Automaattisten tunnistajien suorituskyky

Kuten mainittua, nykyisten automaattisten sanaluokkien tunnistajien tunnistustarkkuus — englanninkielistä kirjakieltä analysoitaessa — 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. Lisäksi useissa tekstilajeissa, kuten uutisiartikkeleissa, lauseiden keskipituus on yli 20 sanaa, jolloin edellämainitullakin tunnistustarkkuudella jokaisessa lauseessa on keskimäärin ainakin yksi virhe (Manning & Schütze, 1999). 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.

Tarkkuustuloksia arvioidessa tulee myös ottaa huomioon varsin korkea lähtötaso: jo yksinkertaisimmalla metodilla, eli valitsemalla kullekin sanalle se sanaluokka, joka esiintyy harjoitusaineistossa useiten annetun sanan yhteydessä, saavutetaan 90% tunnistustarkkuus (Charniak ym., 1993).

On myös mielenkiintoista huomata, että edes ammattilaisten käsin luokittelemat aineistot eivät saavuta täydellistä tunnistamistarkkuutta: ihmisten sanaluokkien tunnistamistarkkuuden on arvioitu olevan noin 97% (Manning, 2011), mikä vastaa edellämainittua automaattisten tunnistajien huipputulosta.

2.4 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.

2.5 Harjoitusaineisto ja sanaluokkasetit

TODO Tässä kappaleessa mainitaan yleisimmät harjoitusaineistot ja sanaluokkasetit (Brown, Penn Treebank). Mahd. syytä täsmentää, että on kyse ohjatusta oppimisesta. Testiaineiston ja harjoitusaineiston erot; mitä seuraa jos liian erilaiset? Ongelmaan vaikuttaa myös tunnistettavien sanaluokkien määrä: mitä enemmän sanaluokkia, sitä suurempi mahdollisuus monitulkintaisuuteen. Harjoitusaineiston koko. Ongelmallista on myös tunnistamisessa käytettävien sanaluokkien (engl. *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) .

Yleensä puhutaan 8:sta tagista, NLP:llä eri vaatimukset: päätemerkit yms.

3 Transformaatiosäännöt

TODO

4 Markovin piilomallit

TODO + yksinkertainen

+ vaatii vähän muistia vrt. tilastolliset menetelmät

+ lopputuloksena saadaan lingvistisesti merkityksellistä ja helposti tulkittavaa dataa

— sääntöjä

- esimerkkikuva Markovin ketjusta hitaampi opettaa

- markov-oletus - markovin ketjut - bigrammit/trigrammit - miksi piilotettu?

huonompi tarkkuus

4 Markovin piilomallit

Markovin malli (mm. Rabiner, 1989) kuvaa sellaista stokastista prosessia, ~~jonka vallitseva~~ joka toteuttaa ns. Markov-ominaisuuden: seuraava tila riippuu aina vain N:stä edeltävist10001000v10001000~~tiloista. Yksinkertaisimmillaan malli koostuu havaintoja kuvaavista tiloista, joille kullekin on~~ stä tilasta. Markov-ominaisuus on siis eräänlainen riippumattomuusol yksinkertaistaa stokastisen prosessin tilan estimointia rajoittamalla tilasiirtymien historian määrä~~tty tilasiirtymien~~ ä. N:nen asteen Markov-ominaisuuden toimiessa esimerkiksi todennäkö~~isyydet~~isyys

$$P(x_k|x_1,\ldots,x_{k-1})$$

voidaan laskea huomattavasti yksinkertaisemmin tarkastelemalla vain N:ää edellistä tilaa:

$$P(x_k|x_{k-N},\ldots,x_{k-1})$$

Markov-ominaisuudesta puhuttaessa on yleensä kyse juuri ensimmäisen asteen Markov-omina jolloin $N = 1$. ~~Usein prosessin tila ei kuitenkaan ole suoraan havaittavissa~~Käytännössä tämä tarkoittaa sitä, ~~eli tila on piilotettu, joskin havainnosta riippuvainen ;~~ että tarkastellaan jotakin kahta peräkkäistä tilaa — nykyistä sekä tulevaa. Markov-ominaisuutta voidaan kuitenkin laajentaa myös korkeampiin asteisiin, jolloin myös tarkasteltavien tilasarjojen pituudet kasvavat. Sanaluokkia tunnistaessa näitä tilasarjoja vastaavat n:n peräkkäisen sanaluokan sarjat, eli ns. n-grammit.

Yksinkertaisimmillaan Markovin mallia voidaan kuvata tilakoneena, joka koostuu havaittavia tapahtumia kuvaavista tiloista, sekä tilasiirtymämatriisista, josta ilmenevät todennäköisyydet siirtyä kustakin tilasta mihin tahansa muuhun tilaan. Kukin tila on siis itsenäinen ja muistiton.

TODO Mahd. selitykset HMM:n viidestä elementistä (**?**) .

TODO: täll10001000hän esimerkkikuva Markovin ketjusta

Markovin mallin hy01000in on kyse Markovin piilomallista. Sanaluokkien tunnistamisongelma voidaan kuvata kyseisen01000dyllisyytt01000 piilomallina: rajoittaa se, että malli ei pysty itsessään mallintamaan prosesseja, joiden tilat eivät ole suoraan havaittavissa. Useimmissa mielenkiintoisissa tapauksissa prosessin tilat eivät kuitenkaan suoraan vastaa havaintoja, eli prosessin tila — vaikkakin havainnosta riippuvainen — on piilotettu: esimerkiksi sanaluokkien tunnistamisongelmassa havainto (sana) on tiedossa, tila (sanaluokka) on riippuvainen havainnosta, mutta tila itsessään ei ole tiedossa. Tällöin Markovin malli tulee laajentaa Markovin piilomalliksi, jossa havainto on aina sitä vastaavan tilan todennäköisyysfunktio.

TODO: tähän kuva Markovin piilomallista

4.1 Lähtökohta

Tunnistamisongelmassa siis haetaan annetulla lauseelle sitä todennäköisimmin vastaavaa sanaluokkien sarjaa. Todennäköisyyttä voidaan mallintaa funktiolla

$$p(w_1, w_2, \ldots, w_n, t_1, t_2, \ldots, t_n)$$

mikä ilmaisee todennäköisyyden sille, että jokin lause $w_1 \ldots w_n$ esiintyy jonkin sanaluokkasarjan $t_1 \ldots t_n$ yhteydessä. Tällöin ratkaistavaksi jää

$$\arg\max_{t_1 \ldots t_n} p(w_1, w_2, \ldots, w_n, t_1, t_2, \ldots, t_n)$$

eli sanaluokkasarja $t_1 \ldots t_n$, jolla saadaan maksimiarvo edeltävästä funktiosta. Mahdollisten sanaluokkasarjojen määrä kuitenkin kasvaa eksponentiaalisesti sanojen ja sanaluokkien määrän mukaan, jolloin maksimiarvon ratkaiseminen on epätehokasta.

4.2 Tunnistusongelma Markovin piilomallina

Markovin piilomallien avulla edellämainittu funktio p voidaan kuvata muodossa

$$p(w_1, w_2, \ldots, w_n, t_1, t_2, \ldots, t_n) = \underbrace{\prod_{i=1}^{n+1} q(t_i|t_{i-2}, t_{i-1})}_{\text{Markovin ketju}} \prod_{i=1}^{n} e(w_i|t_i)$$

missä t_0 ja t_{-1} ovat lauseen alkuun lisättyjä alkusanaluokkia, ja t_{n+1} on päätemerkki-sanaluokka. Mallin ensimmäinen parametri

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

laskee todennäköisyyden sanaluokalle t_i, kun kaksi edeltävää sanaluokkaa ovat t_{i-1} ja t_{i-2}. Parametri voidaan myös mieltää todennäköisyytenä trigrammille t_{i-2}, t_{i-1}, t_i. Tästä ~~trigrammista~~ voidaan päätellä, että kyseessä on ~~ns. toisen~~ kolmannen asteen Markovin piilomalli. Mallin toinen parametri

$$e(w_i|t_{1i})$$

laskee todennäköisyyden sille, että sana w_i esiintyy sanaluokan t_i yhteydessä.

4.2.1 Parametrien estimointi

Yksinkertaisimmillaan parametri $q(t_i|t_{i-2}, t_{i-1})$ voidaan estimoida laskemalla

$$q(t_3|t_1, t_2) = \frac{f(t_1, t_2, t_3)}{f(t_1, t_2)}$$

missä funktio f merkitsee annetun n-grammin lukumäärää harjoitusaineistossa. Brants (2000) kuitenkin osoittaa, ~~ettei~~ että datan harvuuden vuoksi tällainen estimaatti ei ole käyttökelpoinen~~, sillä~~ : laajassakaan harjoitusaineistossa ei ole tarpeeksi montaa kappaletta kutakin eri trigrammia. Lisäksi osa trigrammeista t_i, t_{i+1}, t_{i+2} ovat väistä-

mättä sellaisia, että $f(t_i, t_{i+1}, t_{i+2}) = 0$, jolloin koko sarja $t_1 \dots t_n$ saa todennäköisyyden 0. Luotettavampi tapa estimoida arvoa q on hyödyntää trigrammien lisäksi myös harjoitusaineistosta johdettujen uni- ja ~~digrammien~~ bigrammien suhteellisia frekvenssejä:

$$Unigrammi : P(t_3) = \frac{f(t_3)}{N}$$
$$Digrammi Bigrammi : P(t_3|t_2) = \frac{f(t_2, t_3)}{f(t_2)}$$
$$Trigrammi : P(t_3|t_1, t_2) = \frac{f(t_1, t_2, t_3)}{f(t_1, t_2)}$$

missä N merkitsee harjoitusaineiston sanojen kokonaislukumäärää. Nyt funktion q arvoa voidaan silottaa interpoloimalla edellämainittuja n-grammeja:

$$q(t_3|t_1, t_2) = \lambda_1 P(t_3) + \lambda_2 P(t_3|t_2) + \lambda_3 P(t_3|t_1, t_2)$$

missä $\lambda_1 + \lambda_2 + \lambda_3 = 1$ ja $\lambda_1, \lambda_2, \lambda_3 \geq 0$ (TODO muut silotusmenetelmät, perustelu interpolaatiolle). Vastaavasti todennäköisyys e voidaan estimoida vertaamalla sana-sanaluokka-yhdistelmän frekvenssiä pelkän sanaluokan frekvenssiin:

$$e(w_3|t_3) = \frac{f(w_3, t_3)}{f(t_3)}$$

4.2.2 Tuntemattomien sanojen käsittely

Edellinen todennäköisyyden e estimaatti ei kuitenkaan ole luotettava, jos sana w ei esiinny harjoitusaineistossa kertaakaan. Tällöin $e(w|t) = 0$ millä tahansa sanaluokalla t, ja samoin $p(w_1, w_2, \dots, w_n, t_1, t_2, \dots, t_n) = 0$, jos yksikään sanoista $w_1 \dots w_n$ on tuntematon. Yksinkertaisin ratkaisu on määrätä tuntemattoman sanalle aina harjoitusaineiston yleisin sanaluokka, käytännössä substantiivi. Joitain kieliä — kuten englantia — tulkittaessa voidaan saavuttaa parempia tuloksia suffiksianalyysin (Samuelsson, 1993) avulla. Tällöin tarkoituksena on hyödyntää sitä seikkaa, että sanan

pääte on usein vahva indikaattori sen sanaluokasta. Lisäksi Toutanova ym. (2003) ovat esittäneet, kuinka seuraavassa kappaleessa esiteltäviä log-lineaarisia malleja voidaan hyödyntää tuntemattomien sanojen käsittelyssä.

4.2.3 Viterbin algoritmi

Tässä kappaleessa kuvaillaan lyhyesti Viterbin algoritmia, jolla ratkaistaan tehokkaasti em. arvo $\arg\max_{t_1 \ldots t_n} p(w_1, w_2, \ldots, w_n, t_1, t_2, \ldots, t_n)$.

4.3 Parannukset (tjsp)

Tähän kuvaukset Cyclic Dependency Network-menetelmästä (Toutanova ym., 2003), mahd. ohjaamattomasta oppimisesta (Banko & Moore, 2004). The trigram assumption is arguably quite strong, and linguistically naive. However, it leads to models that are very useful in practice.

5 Log-lineaariset mallit

TODO

6 Menetelmien vertailua/analyysia

TODO Tässä kappaleessa esitellään tunnistimien eri toimintaympäristöt (esim. erilaiset kielet, aineistot, harjoitusaineiston saatavuus), ja pohditaan miten eri menetelmät toimivat eri olosuhteissa; vastataan siis tutkimuskysymykseen (miten eri menetelmät eroavat toisistaan, ja soveltuvatko jotkin menetelmät toisia paremmin tietyille toimintaympäristöille).

Toisaalta voisi olla mielekkäämpää perustella menetelmien etuja jo aikaisemmissa kappaleissa, niiden esittelyiden yhteydessä. Menetelmien esitysjärjestys on sellainen, että myöhempi menetelmä tavallaan vastaa aikaisemman menetelmän puutteisiin. Lisäksi tässä kappaleessa esiteltäviä johtopäätöksiä voisi jättää yhteenvetokappaleeseen; kokonaisen kappaleen verran johtopäätöksiä ja merkityksellistä analyysia vaikuttaa turhan kunnianhimoiselta tavoitteelta. Harkitsen siis tämän kappaleen poistamista.

7 Yhteenveto

TODO Yhteenvedossa kerrataan työn pääkohdat lyhyehkösti johtopäätöksiä tehden. Siinä voi myös esittää pohdintoja siitä, minkälaisia tutkimuksia aiheesta voisi jatkossa tehdä. viittaa kirjallisuuskatsauksen tarkoitukseen ja kertoo ”päätulokset”.

Kirjallisuutta

Banko, M. & Moore, R. C. (2004). *Part of speech tagging in context*. Proceedings of the 20th conference on Computational Linguistics, ACL, s. 556.

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.

Charniak, E., Hendrickson, C., Jacobson, N., & Perkowitz, M. 1993. *Equations for part-of-speech tagging*. Proceedings of AAAI-93, s. 784–789.

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.

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.

Manning, C. D. & ~~Sch01000~~Schütze, H. 1999. *Foundations of statistical natural language processing*. Cambridge, MA~~:~~. MIT Press

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.