Aleksi Pekkala

Sanaluokkien automaattisen tunnistamisen menetelmät

Tietotekniikan kandidaatintutkielma

5. 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: 15+0

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

Avainsanat: luonnollisten kielten käsittely, sanaluokkien tunnistaminen, koneoppiminen

Abstract: Englanninkielinen versio tiivistelmästä.

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

Sisältö

1 JOHDANTO 1
1.1 Tutkimuskysymys 1
1.2 Tutkimusmenetelmä 2

2 SANALUOKKIEN AUTOMAATTINEN TUNNISTAMINEN 3
2.1 Mihin sanaluokkien tunnistamista käytetään? 3
2.2 Sanaluokkien tunnistamisen lyhyt historia 3
2.3 Miksi sanaluokkien tunnistaminen on ongelmallista? 3
2.4 Automaattisten tunnistajien suorituskyky 5
2.5 Sanaluokkien tunnistajan vaatimukset 5

3 SANALUOKKIEN TUNNISTAMISEN MENETELMÄT 7
3.1 Markovin piilomallit 7
3.2 Suurimman entropian periaate 7
3.3 Transformaatiosäännöt 7

4 MENETELMIEN VERTAILUA/ANALYYSIA 8

5 YHTEENVETO 9

KIRJALLISUUTTA 10

1 Johdanto

Sanaluokkien automaattinen tunnistaminen (*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.

1.1 Tutkimuskysymys

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 saavut-

tamat tunnistustarkkuudet liikkuvat kaikki suunnilleen samoissa lukemissa. Kun ilmiselvin valintakriteeri on näin poissuljettu, on tehokkaimman menetelmän valinta 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 toimintaympäristöissä, ja mitkä ovat niiden oleellisimmat vahvuudet sekä heikkoudet. Menetelmien suhteelliset ominaisuudet johdetaan tarkastelemalla lähemmin kunkin menetelmän toimintaa, sekä menetelmään liittyvää tutkimuskirjallisuutta.

1.2 Tutkimusmenetelmä

TODO tutkimusmenetelmän esittely

Tutkielma rakentuu seuraavasti: toisessa luvussa annetaan lyhyt johdanto sanaluokkien tunnistamiseen ja sen haasteisiin. Kolmannessa luvussa tarkastellaan kolmea täysin erilaista tunnistusmenetelmää, ja esitellään niiden keskeiset ominaisuudet. Lopuksi tarkastelemme menetelmien suhteellista tehokkuutta erilaisten kriiterien kautta, pyrkien hahmottamaan kunkin menetelmän vahvuudet.

2 Sanaluokkien automaattinen tunnistaminen

TODO Tässä kappaleessa lyhyt johdanto sanaluokkien tunnistamiseen.

2.1 Mihin sanaluokkien tunnistamista käytetään?

TODO päällekkäisyydet johdannon kanssa

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.

2.2 Sanaluokkien tunnistamisen lyhyt historia

TODO

2.3 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). Ongelmaan vaikuttaa myös tunnistettavien sanaluokkien määrä: mitä enemmän sanaluokkia, sitä suurempi mahdollisuus monitulkintaisuuteen.

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).

2.4 Automaattisten tunnistajien suorituskyky

Kuten mainittua, nykyisten automaattisten sanaluokkien tunnistajien tunnistustarkkuus 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.

2.5 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 kohdistami-

nen uudelle kielelle, aineistolle tai sanaluokkasetille on mahdollisimman vaivatonta.

3 Sanaluokkien tunnistamisen menetelmät

TODO Tässä kappaleessa esitellään kolme yleistä tunnistamismenetelmää; tilastollinen, ominaisuuksiin perustuva sekä transformaatiosääntöihin perustuva. Perustellaan em. luokittelu sekä esiteltävien menetelmien valintaperusteet. Menetelmien esittelyjen pohjat löytyvät tutkimussuunnitelmasta, joskin vielä hieman puutteellisina, siksi niitä ei ole vielä kopioitu tähän.

3.1 Markovin piilomallit

TODO alkuperä, toiminta, suorituskyky ym.

3.2 Suurimman entropian periaate

TODO

3.3 Transformaatiosäännöt

TODO

4 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).

5 Yhteenveto

TODO Yhteenvedossa kerrataan työn pääkohdat lyhyehkösti johtopäätöksiä tehden. Siinä voi myös esittää pohdintoja siitä, minkälaisia tutkimuksia aiheesta voisi jatkossa tehdä.

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.

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. & Schütze, H. 1999. *Foundations of statistical natural language proces-*

sing. 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.