Mielipiteiden louhinta*

Sakari Jokinen

29. lokakuuta 2007

1 Johdanto

Internet ja WWW ovat täynnä erilaisia mielipiteitä esittäviä tekstejä. Sivustot jotka tarjoavat mahdollisuuden kirjoittaa sekä lukea tuotteiden arvosteluja ovat yksi esimerkki. Toisaalta blogit tai muut yleiset sivut voivat myös sisältää mielenkiintoisia mielipiteitä. Mielipiteitä esittävien sivustojen määrä tai ainakin löydettyjen sivustojen määrä kasvaa myös vauhdilla. Peter Turney kertoo etsineensä googlesta hakusanoilla "akumal travel review"[1] vuonna 2002 ja saaneensa noin 5000 osumaa [Tur02]. Tänä päivänä[2] vastaava haku tuottaa noin 202 000 osumaa.

Mielipiteiden louhinnassa (engl. opinion mining) yritetään tunnistaa ja tiivistää tietoa esitetyistä mielipiteistä. Haulla "akumal travel review" luonnollisestikaan jokainen löydetyistä osumistä ei ole oikea arvostelu vaan seassa on myös muunkaltaisia sivuja. Eräs avoin ongelma onkin miten ylipäätään voidaan löytää tuotearvostelut tästä joukosta. Toisaalta kukin löydetty arvostelu kertoo vain yksittäisen ihmisen tai yhteisön mielipiteen. Jos otetaan ensimmäinen googlen antama osuma, niin se on epäilemättä Pagerank algoritmin mukaan merkittävä [BrP98], mutta mutta tämä ei auta meitä saamaan yleiskuvaa vallitsevasta mielipiteestä [YuH03].

Loput tutkielmasta on järjestetty seuraavasti. Ensiksi luvussa 2 määritellään yleisiä käsitteitä, joita tutkielmassa käytetään. Luvussa 3 yritetään esittää menetelmiä joilla voidaan löytää mielipiteitä. Mielipide on usein asennoitunut joko positiivisesti tai negatiivisesti. Luvussa 4 yritetään tunnistaa tätä mielipiteiden asennoitumista. Luvussa 5 käsitellään mielipiteen lähteen tunnistamista. Luvussa 6 käsitellään mielipiteen kohteen tunnistamista tekstiaineistosta. Vertailujen tunnistamista käsitellään luvussa 7. Mielipiteiden takana on usein perusteluja tai syitä. Luvussa 8 esitellään yksi menetelmä näiden syiden löytämiseksi. Luku 9 käsittelee mielipidelouhinnan haasteita. Viimeisenä on yhteenveto koko tutkielmasta (luku 10).

*Seminaari: Tiedon louhiminen WWW:stä, Helsingin yliopisto, Syksy 2007

[1] Akumal on pieni kylä Meksikossa

[2] 25.10.2007

2 Käsitteitä

2.1 Materiaali

Internetistä löytyvät tekstit ovat tyypiltään hyvinkin erilaisia. Liu, Hu ja Cheng tunnistavat kolme erityyppistä tuotearvostelua [LHC05]:

1. Puolesta ja vastaan
2. Vapaateksti
3. Yhdistetty

Puolesta ja vastaan-arvostelussa arvostelija kirjoittaa lyhyitä lauseita kustakin tuotteen hyvästä ja huonosta puolesta. *Vapaateksti-arvostelussa* arvostelija kirjoittaa vapaasti tuotteesta. Näiden yhdistelmässä arvostelija erittelee yksittäisiä puolesta ja vastaan piirteitä, sekä kirjoittaaa vapaasti muotoillun arvostelun.

Kirjoitetun tekstin tekstuaaliset osat ovat myös tärkeitä mieliteiden louhinnassa. *Fraasi* (engl. phrase) on lyhyt osa tekstistä. Tyypillisesti fraasit ovat osia lauseista.

2.2 Subjektiivisuus

Lause tai yleisemmin väite on *subjektiivinen*, jos se esittää jonkin mielipiteen. Vastakohtana subjektiivisille väitteille on *objektiivinen* väite joka esittää tosiasian. Fraasi "kala on hyvää" esittää mielipiteen kalasta ja on siten subjektiivinen.

Kim ja Hovy erittelevät neljä eri piirrettä jotka mielipiteellä voi olla: mielipiteen lähde, mielipiteen kohde, väite sekä asenne (engl. sentiment)[KiH04]. Mielipide tarkoittaa tällöin sitä, että mielipiteen lähde uskoo väitteen kohteesta ja useassa tapauksessa liittää jonkin asenteen tähän uskomukseen.

Mielipiteen *asenneorientaatio* (engl. semantic orientation) määrittää mihin kohtaan mielipiteen asenne asettuu akselilla hyvä/neutraali/huono[HaW00].

Tekstissä mielipide ilmenee käytännössä lauseena tai fraasina. Voidaan myöskin puhua kokonaisen tekstin esittämästä mielipiteestä.

2.3 Tuloksista

Oletetaan, että joukko O sisältää kaikki oliot, R_α ($R_\alpha \subseteq O$) sisältää kaikki luokkaan α kuuluvat oliot ja Q_α ($Q_\alpha \subseteq O$) sisältää kaikki oliot jotka luokittelualgoritmi määrittelee kuuluvaksi luokkaan α. Luokittelualgoritmin *palautus* (engl. recall) P_α on suhde oikein luokiteltujen ja kaikkien luokkaan kuuluvien olioiden kesken (1) [Ni07]. Jos luokka α on selvä asiayhteydestä niin se voidaan jättää merkitsemättä.

$$P_\alpha = \frac{|R_\alpha \cap Q_\alpha|}{|R_\alpha|} \tag{1}$$

Tarkkuus (engl. precision) T_α on luokkaan oikein luokiteltujen olioiden osuus (2) [Ni07].

$$T_\alpha = \frac{|R_\alpha \cap Q_\alpha|}{|Q_\alpha|}. \tag{2}$$

Yksinkertaisesti luokittelemalla kaikki oliot luokkaan α kuuluviksi saadaan $P_\alpha = 1$, mutta T_α on silloin mahdollisesti hyvinkin matala. Koska tarkkuus ja palautus liittyvät tällä tavalla yhteen niin usein käytetään näiden *harmonista keskiarvoa* (engl. harmonic mean) F_α (3) [Ni07].

$$F_\alpha = \frac{2 * P_\alpha * T_\alpha}{P_\alpha + T_\alpha} \tag{3}$$

Luokittelualgoritmin *virheettömyys* (engl. accuracy) V on toinen tapa tutkia luokittelun laatua. Virheettömyys on kaikkien luokittelijan oikein luokittelemien olioiden osuus koko oliojoukosta (4).

$$V = \frac{\sum_k |R_k \cap Q_k|}{|O|} \tag{4}$$

Ni ja muut [Ni07] ovat yksi harvoista, jotka määrittelevät miten P_α tai T_α lasketaan. Tässä tutkielmassa on oletettu, että muidenkin artikkeleiden käyttämät palautus ja tarkkuus vastaavat heidän määritelmiään. Tämä ei ole välttämättä selvää. Esimerkiksi Kobayashi ja Takeda määrittelevät tarkkuuden kaavalla $\frac{|R_\alpha|}{|Q_\alpha|}$[KoT00]. Toisaalta heidän määrittelyssään voi olla painovirhe. Tällä tavalla määriteltynä tarkkuuden arvo ei pysy välillä $[0 \ldots 1]$.

Kirjallisuudessa tuloksia on annettu eri tarkkuuksilla. Tässä tutkielmassa ensiksi kaikki prosentteina annetut tulokset on muunnettu välille $[0 \ldots 1]$, jonka jälkeen kaikki tulokset ovat pyöristetty sadasosan tarkkuudelle.

3 Subjektiivisuusluokittelu

Yksinkertainen menetelmä lauseen subjektiivisuuden määrittämiseen on laskea yhteen lauseen sanojen asenneorientaatiota [HaW00]. Yu ja Hatzivassiloglou käyttävät sanojen asenneorientaatioita piirteinä luokittelussa [YuH03]. Kokonaisen dokumentin tasolla heidän menetelmän saavutti 0,97 F arvon[3]. Lausetasolla heidän menetelmänsä tarkkuus oli 86% (palautus 91%).

Mahdollinen käyttökohde subjektiivisuusluokittelulla on tekstien laadun määrittely. Esimerkiksi laadukkaiden blogien tai hyödyllisten tuotearvosteluiden tunnistaminen ovat kiinnostavia alueita. Subjektiivisuus kertoo ainakin jotain tekstistä. Ghose ja Ipeirotis tutkivat tuotearvosteluiden subjektivisuuden vaikutusta toisaalta tuotteen menekkiin ja toisaalta

[3]Paperissa ei anneta erikseen arvoja palautuksella ja tarkkuudelle

tuotearvostelun hyödyllisyyteen kuluttajan kannalta [GhI07]. He toteavat, että arvosteluilla joissa on sekoitus objektiivisia ja hyvin subjektiivisia lauseita on positiivinen vaikutus tuotteen myyntiin. Mitä objektiivisempi tuotearvostelu on, sitä hyödyttömämmäksi kuluttajat tulkitsevat sen. Toisaalta Zhang ja muut eivät havainneet subjektivisuudella tälläistä vaikutusta [ZhV06]. He totesivat parhaimman ennustajan tuotearvostelun hyödyllisyydelle olevan tekstin kielellinen tyyli. Zhang ja muut määrittävät tekstin subjektivisuuden laskemalla tekstistä löytyviä subjektiivisia sanoja. Ghose ja Ipeirotis käyttivät koulutettavaa luokittelijaa. Ero tuloksissa johtuu mahdollisesti eri tavasta tunnistaa subjektiivisuus.

4 Asenneorientaatioluokittelu

4.1 Sanojen asenneorientaatio

Sanoilla kuten "kaunis" ja "hyvä", on positiivinen orientaatio kun taas sanoilla "ruma" tai "huono" on negatiivinen orientaatio. Toisaalta sanoilla kuten "punainen" tai "iso" on tyypillisesti neutraali orientaatio.

Miten tuntemattoman sanan asenneorientaatio voidaan määrittää? Useat menetelmät sanojen asenneorientaation määrittämiseen perustuvat pienen käsin määritellyn siemensanaston laajentamiseen automaattisesti [YuH03, HuL04, KiH04, HaM97]. Tässä voidaan käyttää jo olemassaolevia sanastoja joissa sanoille on määritelty vastakohtia tai synonyymeja [HuL04, KiH04] . Esimerkiksi jos sanalle "hauras" on määritelty negatiivinen orientaatio niin sen synonyymille "heikko" voidaan myös antaa negatiivinen orientaatio. Vastaavasti vastakohdalle "vahva" voidaan antaa positiivinen orientaatio.

Fraaseissa "reilu ja oikeudenmukainen" ja "brutaali ja korruptoitunut" konjunktio yhdistää sanoja joilla on sama orientaatio. Toisaalta "reilu, mutta brutaali" on lause, jossa konjuntio erottaa eri suuntiin orientoituneita sanoja. Jos tiedetään, sanan "reilu" orientaatio, niin voidaan päätellä, että "oikeudenmukainen" on positiivisesti orientoitunut ja vastaavasti, että "brutaali"on negatiivisesti orientoitunet. Tästä voidaan päätellä, että "korruptoitunut" on myös negatiivisesti orientoitunut. Konjunktioissa esiintyvät sanat muodostavat verkon solmut, jossa kaaret solmujen välillä joko kääntävät tai säilyttävät orientaation. Hatzivassiloglou ja McKeown toteavat, että parhaimmillaan näin saadaan määritelty oikea orientaatio 83% sanoista [HaM97].

Kaikkien kirjoitettujen sanojen merkitystä ei kuitenkaan voida selvittää. Homonyymit ovat sanoja jotka kirjoitetaan samalla tavalla, mutta joiden tarkoitus on eri. Tekstistä löydetyn homonyymin asenteen määrittämiseksi tarvitaan sanan kontekstia[HaM97].

4.2 Dokumentin asenneorientaatio

Yksinkertainen menetelmä tekstin asenneorientaation määrittelyyn olisi laskea tekstissä esiintyvät asenneorientoituneet sanat. Sanat"kaunis" ja "hyvä" ovat mahdollisesti riippumatta lauseyhteydestä positiivisesti latautuneita sanoja. Kaikilla sanoilla näin ei kuitenkaan ole [Tur02]. Puhuttaessa auton ohjaamisesta väitteessä "ennalta-arvaamaton ohjaus" sanalla "ennalta-arvaamaton" on selkeästi negatiivinen orientaatio. Puhuttaessa elokuvan juonesta väitteessä "ennalta-arvaamaton juoni" sanalla "ennalta-arvaamaton" on positiivinen orientaatio.

Koska sanojen asenneorientaatio riippuu siitä missä yhteydessä niitä käytetään, niin eräs ratkaisu on yrittää yksittäisten sanojen sijasta pyrkiä määrittämään tekstin osien asenneorientaatiota. Mikä tahansa fraasi tekstistä ei kuitenkaan ole mielekäs tutkittavaksi. Fraasista "ja ennalta-arvaamaton" ei voida sanoa oikestaan mitään kun taas fraasilla "ennalta-arvaamaton juoni" on selvästä asenneorientaatio.

Eräs tapa mielekkäiden fraasien eristämiseen on valita vain fraaseja jotka vastaavat jotain annettua sanaluokkiin perustuvaa hahmoja [Tur02]. Esimerkiksi hahmo joka koostuu sanaluokkamerkeistä $< adjektiivi >< substantiivi >$ kerää kaikki adjektiivista ja substantiivistä koostuvat kahden sanan mittaiset fraasit tektistä. Turney tutki arvostelutekstien luokittelua käyttäen näitä fraaseja [Tur02]. Fraasin asenneorientaation määrittämiseksi Turney käyttää fraasin ja termin "excellent" tai "poor" todennäköisyyttä esiintyä yhdessä. Käytännössä Turney etsii AltaVista hakukoneen *NEAR* operaattoril-

Taulukko 1: Korrelaatio merkitsee asenneorientaation arvon ja arvostelun arvosanan korrelaatiota

Aihepiiri	V	Korrelaatio
Autot	0,84	0,46
Pankit	0,80	0,62
Elokuvat	0,66	0,36
Matkailukohteet	0,79	0,41

la sivuja joissa fraasi ja jompikumpi näistä sanoista esiintyi. Löydettyjen osumien määrä sanan "excellent" lähellä määrittää kuinka positiivisesti fraasi on orientoitunut ja vastaavasti sanan "poor" kanssa kuinka negatiivisesti fraasi on orientoitunut. Koko tekstin asenneorientaatio on tekstistä eristettyjen fraasien asenneorientaatioiden keskiarvo.

Taulukossa 1 on verrattu tekstille laskettua asenneorientaatiota arvosteluihin joissa on arvosana (1-5) sekä binäärinen arvostelu "suositeltu" tai "ei suositeltu". Lähes kaikilla aihepiireillä niin tarkkuus kuin korrelaatiokin ovat korkeita. Turney selittää elokuva-arvosteluiden matalan tarkkuuden sillä, että elokuva-arvostelut viittavat elokuvien sisältämiin epämiellyttäviin tapahtumiin. Luokittelussa nämä viittaukset hyvässäkin elokuvassa tulkitaan negatiivisiksi mielipiteiksi.

4.3 Lauseen asenneorientaatio

Lause on asenneorientaatioltaan positiivinen, negatiivinen tai neutraali riippuen siitä esittääkö se toivottavan tai epätoivottovan mielipiteen.

Lauseen subjektiivisuuden määrittelemiseksi yksinkertainen sanojen asenneorientaation perustuva menetelmä voi olla riittävä [HaW00]. Lauseen asenneorientaation määrittämiseksi tämä ei kuitenkaan ole kovin hyvä menetelmä.

Yksittäisten fraasien asenneorientaatioita ei voi helposti yhdistää koko lausen asenneorientaatioksi[KiH04]. Lauseilla "en usko, että tuote on hyvä" on huomattavan erilainen asenneorientaation kuin "uskon, että tuote on hyvä".

Hu ja Liu summaavat ominaisuussanoja sisältävien lauseiden mielipidesanojen asenneorientaatiot (+1/−1) yhteen lauseen asenneorientaatioksi [HuL04]. He huomioivat kuitenkin lauseet kuten "..., mutta *lause*" jotka voivat vaihtaa koko lauseen orientaation sekä lähellä mielipidesanoja esiintyvät negaatiot kuten "ei".

Taulukko 2: Asenneorientaation luokittelu.

Työkalu	Materiaali	T	P	V
Sentiment Miner [YiN05]	Arvostelut	0,87	0,56	0,86
	WWW sivut	N/A	0,86	0,90
			-	-
			0,91	0,93
Hu ja Liu [HuL04]	Arvostelut	N/A	N/A	0,84

Yi ja Niblack käyttävät luonnollisen kielen tulkintaa analysoitavien lauseiden tulkintaan [YiN05]. Heidän työkalunsa *Sentiment Miner* yhdistää tulkittuja lauseita ennaltamääriteltyihin hahmoihin lauseiden merkityksen määrittämiseen. Hahmo koostuu kolmesta osasta: *verbistä*, *kategoriasta* sekä *kohteesta*. Kohde määrittää sanan lauseessa johon mielipide kohdistuu. Joillain verbeillä, kuten "rakastaa" on oma asenneorientaationsa jolloin kategoriana on + tai − verbin orientaation mukaan. Toisaalta kategoria voi määritellä sanan lauseesta joka määrää lauseen annettavan asenneorientaation.

Luonnollisen kielen tulkinnalla päästään hyviin tuloksiin lauseiden asenneorientaation määrittelyssä (Taulukko 2). T ja P arvot taulukossa 2 ovat materiaalille, jossa on mukana vain asenneorientoituneita lauseita. V sisältää myös neutraaleja tekstejä. Sentiment Minerillä on suurempi tarkkuus kuin Hun ja Liun menetelmällä mikä on oletettavaa koska Sentiment Miner suorittaa kattavamman analyysin luokiteltaville lauseille. Toisaalta Hu ja Liu pääsevät melko lähelle yksinkertaisella tekstin analyysillä.

5 Lähteiden tunnistus

Mielipiteen lähteen määrittämiseksi voidaan olettaa, että teksti sisältää vain tekstin kirjoittajan mielipiteitä. Elokuva-arvostelun kirjoittaja esittää tekstissään vain omia mielipiteitään. Tällöin sivulla tai

sen metadatassa yleensä kerrotaan kuka on arvostelun kirjoittaja. Toisaalta tekstissä voidaan kertoa myös muiden tahojen mielipiteistä [Cho05]. Tekstissä voidaan esimerkiksi kertoa, että "*jonkin* mielestä kala on hyvää". Mielipiteen lähteen määrittely tekstin kirjoittajaksi olisi virheellistä.

Kohdefraasi (engl. topic phrases) on fraasi jossa esiintyy jokin mahdollinen mielipiteen kohde. Lauseessa voi olla useita mahdollisia mielipidelähteitä. kuten esimerkiksi kaikki substantiivit. Kim ja Hovy valitsevat näistä mahdollisista mielipidelähteistä sen joka on lähimpänä lauseesta löydettyä kohdefraasia [KiH04]. Kyseessä ei kuitenkaan ole välttämättä oikea mielipiteen esitys. Kim ja Hovy yrittävät tunnistaa onko kyseessä oikeasti mielipide määrittelemällä kohdefraasin ja valitun mahdollisen mielipidelähteen lähettyvillä olevien sanojen asenneorientaation [KiH04].

Lähimmän lähdekandidaatin valitseminen ei kuitenkaan ole kovin tarkka menetelmä. Mielipiteitä voidaan esittää myöskin toisista mielipiteistä. Jotta määrittely olisi tarkka, pitäisi havaita mikä lauseessa on *välitön lähde* ja mikä *välillinen lähde*, eli lähde joka puhuu allaolevasta mielipiteestä [Cho05]. Lauseessa "raportin mukaan kalastajien mielestä kala on hyvää" on kaksi lähdettä. Kalastajat ovat välittömänä lähteenä mielipiteessä kun taas "raportti" on välillinen lähde joka puhuu kalastajien makutottumuksista. Välillisten lähteiden tunnistamiseen tarvitaan analysoitavan tekstin syvempää ymmärtämistä.

Choi ja muut käyttävät tekstin tekstin rakenteeseen ja merkitykseen perustuvia hahmoja mielipiteen lähteen tunnistamiseen[Cho05]. Kaikkien mahdollisten tunnistettujen mielipiteen lähteiden olettaminen mielipiteen lähteeksi ei ole kovin tarkkaa (T = 0,28 P = 0,71). Choi ja muiden menetelmä pärjää huomattavan paljon paremmin (T = 0,79 P = 0,59).

6 Kohteiden tunnistus

Johonkin tiettyyn kohteeseen kohdistuvia mielipiteitä on suhteellisen helppoa löytää. Tähän voidaan käyttää esimerkiksi tavallista hakukonetta. Mitä voidaan tehdä jos emme etukäteen tiedä mielipiteen kohdetta? Toisaalta mielipiteen kohteena voi olla olion jokin *ominaisuus* (engl. feature) koko olion sijasta. Vaikka tietäisimme mistä olioista olemme kiinnostuneita, niin emme välttämättä tiedä minkälaisia ominaisuuksia teksti väittää näillä olioilla olevan.

Mielipidettä vastaavassa fraasissa on *eksplisiittinen kohde* jos teksti mainitsee kohteen. Aina näin ei kuitenkaan ole. Esimerkiksi lauseessa "kamera ei helposti mahdu taskuun" esitetään mielipide kameran koosta, mutta kohdetta "koko" ei mainita lauseessa. Kohde "koko" on *implisiittinen kohde*.

6.1 Eksplisiittiset kohteet

Hu ja Liu tutkivat tuotteiden ominaisuuksien tunnistamista internetistä löytyvien tuotearviointien yhteydessä [HuL04b]. *Yleinen fraasijoukko* (engl. frequent phrase) on joukko tekstissä usein yhdessä esiintyviä sanoja. Hu ja Liu käyttävät yleisiä fraasejoukkoja mahdollisina ominaisuuksina. Kuten taulukosta 3 nähdään kaikki yleiset fraasijoukot eivät ole oikeita ominaisuuksia.

Yleisissä fraasijoukoissa ei oteta huomioon sanojen järjestystä [HuL04b]. Toisaalta tekstissä sanojen järjestyksellä ja läheisyydellä on merkitys.

Oletetaan, että yleinen fraasijoukko f koostuu n sanasta $w_1, w_2, \ldots, w_n$ ja funktio $d(s, w, w')$ laskee sanojen w ja w' etäisyyden lauseessa s.

Yleinen fraasijoukko f on *kompakti* lauseessa s, jos kaikilla sanapareilla w_i ja w_{i+1} $d(s, w_i, w_{i+1}) < 4$. Jos f on kompakti ainakin kahdessa lauseessa niin se on *kompakti fraasijoukko* (engl. compact phrase).

Yleisen fraasijoukon f *p-tuki* (engl. p-support) on niiden lauseiden määrä joissa f esiintyy substantiivina ilman, että f sisältyisi lauseessa johonkin toiseen yleiseen fraasijoukkoon.

Karsimalla fraasijoukkoja näillä kahdella kriteerillä tarkkuus saadaan nostettua 72 prosenttiin (taulu 3).

Yleiset fraasitjoukot eivät tunnista harvoin lauseissa esiintyviä ominaisuuksia. Hu ja Liu käyttävät ominaisuuksien tunnistamiseen toista menetelmää [HuL04b]. Ominaisuudet esiintyvät usein asenneorientoituneiden sanojen yhteydessä. Jos lauseesta ei löydy yleisen fraasijoukon tunnistamaa ominaisuutta niin heidän tunnistamaansa asenneorientoitunutta sanaa lähinnä oleva subjekti tulkitaan ominaisuu-

Taulukko 3: Ominaisuuksien tunnistaminen tekstistä

Menetelmä	P	T
Pelkät yleiset fraasitjoukot	0,68	0,56
Kompaktit fraasit sekä p-tuki > 3	0,67	0,72
Harvoin esiintyvät ominaisuudet	0,80	0,72

Taulukko 4: Ominaisuuksien tunnistaminen puolesta ja vastaan lauseista

Menetelmä	Puolesta		Vastaan	
	P	T	P	T
Yleiset fraasijoukot	0,44	0,51	0,47	0,48
Subjektit	0,65	0,60	0,70	0,36
Assosiaatiosäännöt	0,90	0,89	0,82	0,79

deksi. Kuten taulusta 3 näkyy tämä viimeinen menetelmä nostaa palautusta huomattavasti.

Liu, Hu ja Cheng käyttävät assosiaatiosääntöjen louhintaa muodostaakseen sääntöjä puolesta ja vastaan-lauseista. Säännössä $p_1, \ldots, p_i \rightarrow [ominaisuus]\ p_i, \ldots, p_i$ merkitsevät puheenosia (engl. part of speech) kuten adjektiivejä tai substantiivejä [LHC05]. He merkitsevät käsin [*ominaisuus*] merkillä sanat jotka ovat oikeita ominaisuuksia tekstissä.

Taulukosta 4 nähdään, että tällä menetelmällä saadaan selvästi parempia tuloksia kuin edelläkuvatussa yleisiin fraasijoukkoihin perustuvassa menetelmässä samalla aineistolla [LHC05]. Jopa pelkästään kaikkien lauseiden subjektien olettaminen ominaisuuksiksi pärjää puolesta ja vastaan lauseissa paremmin kuin yleiset fraasijoukot. Ominaisuuksien tunnistamiseen vaikuttaa siis merkittävästi tekstin laatu. Ei ole selvää kuinka hyvin Liun ja muiden assosiaatisäännöt pärjäävät jos kohteena on yleinen teksti. Toisaalta Liu, Hu ja Cheng joutuvat merkitsemään koulutusmateriaalin käsin, mitä ei yleisillä fraasijoukoilla tarvitse tehdä.

6.2 Implisiittiset kohteet

Zhuang ja muut käyttävät hyväkseen kahta yksinkertaista piirrettä elokuva-arvosteluissa[ZJZ06]. Jos

Taulukko 5: Vertailulauseiden tunnistus eri tekstityypeistä [JiL06]

Tekstityyppi	T	P
Arvostelut	0,84	0,80
Uutiseartikkelit	0,75	0,80
Foorumit	0,73	0,83

lause on korkeintaan kolmen sanan mittainen ja se esiintyy joko arvostelun alussa tai lopussa, niin se yleensä kertoo yleisarvion elokuvasta. Toisaalta joihinkin kohteisiin viitataan jollain etukäteen tunnetuilla sanonnoilla. Lause “Hyvin näytelty” viittaa näyttelijöihin kun taas “pakko katsoa” viittaa elokuvaan.

7 Vertailu

Mielipide voi sisältää vertauksen eri tahojen välillä. *Vertaileva lause* esittää kahden tai usemman kohteen välisen suhteen [JiL06]. Vertailevasta lauseesta voidaan mahdollisesti tunnistaa mielipide jotain kohdetta kohtaan, mutta itse vertauksen “X on parempi kuin Y” eristäminen ei onnistu aikaisemmin kuvatuilla menetelmillä.

Jotkin vertaukset on helppo tunnistaa. Helppoja tapauksia ovat esimerkiksi komparatiivit [JiL06]. Toisaalta kaikki vertailut eivät käytä komparatiivejä.

Jindal ja Liu tunnistavat kaksi vaihetta vertauksien louhinnasssa: vertauslauseiden tunnistaminen sekä vertaussuhteiden eristäminen. Jindal ja Liu keskittyvät näistä ensimmäiseen [JiL06]. Taulukosta 5 nähdään, että ainakin vertauslauseiden tunnistaminen on mahdollista. Vertaussuhteiden eristämisestä ei löytynyt kirjallisuutta.

8 Mielipiteen syy

Edellä on käsitelty mielipiteiden eristämistä tekstistä. Toisaalta se syy miksi jokin mielipide on esitetty voi myöskin olla mielenkiintoinen. Esimerkiksi lause “En pidä kamerasta koska sen akku kestää vain

puoli tuntia" sisältää mielipiteen "en pidä kamerasta", mutta se sisältää myös mielenkiintoisen tiedon siitä miksi mielipide on mikä se on. Kameran akku on riittämätön.

Annettavaa yleiskuvaa voidaan tarkentaa käyttämällä asenneorientaatioluokittelun kohteena tekstin lauseita [DLP03]. Tällöin käyttäjä voi nähdä, minkälaiset kohdetta käsittelevät lauseet ovat negatiivisia tai positiivisia. Näin ei kuitenkaan voida eristää tarkasti varsinaisia syitä mielipiteelle.

Kim ja Hovy yrittävät vastata kysymykseen "Mitkä ovat syyt siihen, että arvostelun kirjoittaja pitää tai ei pidä tuotteesta"[KiH06]. Mielipiteiden syiden merkitseminen käsin koulutusmateriaaliin on työlästä. Kim ja Hovy käyttävät hyväkseen yhdistettyjä arvosteluja (kappale **??**). Arvostelussa on erikseen puolesta ja vastaan syitä. Toisaalta vapaatekstissä todennäköisesti esitetään nämä samat syyt. Kim ja Hovy yhdistävät nämä kohdat tekstistä automaattisesti ja käyttävät näin merkittyä tekstiä koulutusmateriaalina.

Kim ja Hovy raportoivat syiden tunnistamisessa parhaimmillaan 0,71 tarkkuuden ja 0,76 palautuksen. Syiden orientaation selvittäminen on vaikeampi ongelma (parhaimmillaan 0,62 tarkkuus ja 0,68 palautus) [KiH06].

9 Mielipidelouhinnan haasteita

Tekstin ymmärtämistä varten kaikki käsitellyt artikkelit käyttävät jotain yhdistelmää luonnollisen kielen käsittelystä sekä jostain koneoppimis -tai luokittelualgoritmista. Luonnollisen kielen käsittely on itsessään hankala ongelma. Toisaalta myös hyvän luokittelijan opettaminen on hankalaa, koska tarvitaan sopiva valmiiksi luokiteltu opetusmateriaali sekä siksi, kuten useassa käsitellyssä paperissa todetaan, että luokittelu ei ole välttämättä ylipäätään mitenkään selvä.

Eri ihmiset voivat luokitella aineiston eri tavoilla [KiH06, ZJZ06]. Opetusmateriaalin muodostamisen ongelma voidaan ratkaista käyttämällä valmiiksi arvioituja tekstejä esimerkiksi sivustoilta joiden formaattiin kuuluu tekstiarvion lisäksi annettava arvosana [DLP03, JiL06, Tur02]. Tämä ei ole kuitenkaan aivan varma menetelmä; eri arvostelijoiden arviointikriteerit voivat olla erilaisia, arvostelut voivat olla epäselvästi kirjoitettuja, arvostelut voivat olla hyvin lyhyitä ja viimeiseksi arvosteluiden jakauma on yleensä epätasainen [DLP03].

Mielipidelouhinnan toimivuus voi riippua myöskin aihepiiristä, jota analysoitava teksti käsittelee[KiH06, DLP03, ZJZ06, Tur02]. Esimerkiksi arviot elektroniikkatuotteista eroavat kieleltään ravintola-arviosta [KiH06]. Siinä missä eloktroniikkatuotteilla on selvä joukko ominaisuuksia, joita arvostelut käsittelevät niin tyypilliset arvostelut ravintoloista sisältävät mielipiteitä abstrakteista sekä vaihtelevista ominaisuuksista. Toisaalta kirjoissa, elokuvissa ja musiikissa voidaan usein käyttää mielipiteenomaisia sanontoja esimerkiksi juonilyhennelmissä, jolloin analyysi voi virheellistä tulkita nämä mielipiteiksi itse elokuvasta tai kirjasta[DLP03].

Kuinka laskennallisesti vaativaa mielipiteiden kerääminen ja luokittelu on? Tähän kysymykseen ei yksikään kirjoittaja vastaa. Esiteltyjen menetelmien käytännöllisyys jää kysymysmerkiksi tältä osin.

Tässä tutkielmassa ei ole käsitelty kysymystä siitä miten louhittava materiaali kootaan. Kuinka internetistä voidaan löytää mielipiteiden louhinnan kannalta mielenkiintoisia tekstejä? Jos haluamme louhia mielipiteitä joiltain ennaltamäärätyiltä sivuilta tämä ei ole tietenkään suuri ongelma. Sivujen rakenne on yleensä tarpeeksi säännöllinen, jotta niiltä voidaan eristää louhinnassa tarvittava teksti.

Yleisessä tapauksessa, jossa kiinnostuksen kohteena ovat kaikki sivustot tehtävä on huomattavasti hankalampi. Luonnollinen ensimmäinen askel on etsiä tavallisella hakukoneella mielenkiinnon kohteeseen liittyviä termejä [DLP03].

Sen jälkeen kun joukko sivuja on löydetty niin niitä voi vielä esiprosessoida poistamalla liian lyhyet tai pitkät lauseet tai kappaleet joissa ei mainita mielenkiinnon kohdetta [DLP03]. Jos halutaan etsiä erityisesti tuotearvosteluita niin voidaan karsia pois esimerkiksi sivut joiden otsikossa ei esiinny sanaa "review" [DLP03]. Sivuja voidaan esiprosessoida etsimällä sivuilta tärkeitä osa-alueita ja käyttää näillä tärkeillä alueilla olevaa tekstiä mielipidelouhinnassa[Son04]. Toisaalta varsinaisia argumentteja siitä kuinka tehokkaita tälläiset esipro-

sesessointiaskeleet ovat ei tutkitussa kirjallisuudessa mainittu.

10 Yhteenveto

Mielipiteiden louhintaa voidaan käyttää mielipiteen kohteen, lähteen tai asenneorientaation määrittämiseen.

Subjektiivisuusluokittelulla pystytään melko luotettavasti tunnistamaan kuinka subjektiivinen jokin lause tai teksti on. Asenneorientaation tunnistaminen on vaikeampaa, mutta siihenkin on menetelmiä. Mielipiteen syyn samoin kuin erilaisten vertailujen tunnistaminen on uusi alue josta on vähän kirjallisuutta. Vertailevia lauseita pystytään tunnistamaan, mutta kirjallisuudessa ei esiinny menetelmiä vertailusuhteiden eristämiseen.

Subjektiivisuuden, lähteiden ja kohteiden tunnistus pelaavat yhteen asenneorientaation sekä vertailujen tunnistamisen kanssa. Etsimällä subjektiivisia lauseita voidaan supistaa aluetta josta etsitään kohteita tai lähteitä. Toisaalta tunnistamalla mahdollisia mielipiteiden lähteitä tai kohteita voidaan myös karsia analysoitavia lauseita. Jos lauseella on asenneorientaatio, niin siinä melko varmasti esiintyy mielipiteen esittäjä ja kohde myöskin.

Mielipiteiden louhinnassa yksi suuri haaste on se, ettei tutkittavaa materiaalia ole helppo ihmistenkään luokitella. Toisaalta erilaisilla tekstityypeillä ja aihepiireillä on suuri vaikutus eri menetelmien toimivuuteen.

Viitteet

[BrP98] Brin, S. ja Page, L., The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30,1–7(1998), sivut 107–117.

[Cho05] Choi, Y., Cardie, C., Riloff, E. ja Patwardhan, S., Identifying sources of opinions with conditional random fields and extraction patterns. *HLT '05: Proc. of the conf. on Human Language Technology and Empirical Methods in Natural Language Processing*, Morristown, NJ, USA, 2005, Association for Computational Linguistics, sivut 355–362.

[DLP03] Dave, K., Lawrence, S. ja Pennock, D. M., Mining the peanut gallery: opinion extraction and semantic classification of product reviews. *WWW '03: Proc. of the 12th int. conf. on World Wide Web*, New York, NY, USA, 2003, ACM Press, sivut 519–528.

[GhI07] Ghose, A. ja Ipeirotis, P. G., Designing novel review ranking systems: predicting the usefulness and impact of reviews. *ICEC '07: Proc. of the 9th int. conf. on Electronic commerce*, New York, NY, USA, 2007, ACM Press, sivut 303–310.

[HuL04] Hu, M. ja Liu, B., Mining and summarizing customer reviews. *KDD '04: Proc. of the 10th ACM SIGKDD int. conf. on Knowledge discovery and data mining*, New York, NY, USA, 2004, ACM Press, sivut 168–177.

[HuL04b] Hu, M. ja Liu, B., Mining opinion features in customer reviews. *Proc. of the 19th nat. conf. on Artificial Intelligence*, Menlo Park, CA, USA, 2004, AAAI Press.

[HaM97] Hatzivassiloglou, V. ja McKeown, K. R., Predicting the semantic orientation of adjectives. *Proc. of the 8th conf. on European chapter of the Assoc. for Computational Linguistics*, Morristown, NJ, USA, 1997, Association for Computational Linguistics, sivut 174–181.

[HaW00] Hatzivassiloglou, V. ja Wiebe, J. M., Effects of adjective orientation and gradability on sentence subjectivity. *Proc. of the 18th conf. on Computational linguistics*, Morristown, NJ, USA, 2000, Association for Computational Linguistics, sivut 299–305.

[JiL06] Jindal, N. ja Liu, B., Identifying comparative sentences in text documents. *SIGIR '06: Proc. of the 29th annual int. ACM SIGIR conf. on Research and development in information retrieval*, New York, NY, USA, 2006, ACM Press, sivut 244–251.

[KiH04] Kim, S.-M. ja Hovy, E., Determining the sentiment of opinions. *COLING '04: Proc. of the 20th int. conf. on Computational Linguistics*, Morristown, NJ, USA, 2004, Association for Computational Linguistics, sivu 1367.

[KiH06] Kim, S.-M. ja Hovy, E., Automatic identification of pro and con reasons in online reviews. *Proc. of the COLING/ACL on Main conf. poster sessions*, Morristown, NJ, USA, 2006, Association for Computational Linguistics, sivut 483–490.

[KoT00] Kobayashi, M. ja Takeda, K., Information retrieval on the web. *ACM Comput. Surv.*, 32,2(2000), sivut 144–173.

[LHC05] Liu, B., Hu, M. ja Cheng, J., Opinion observer: analyzing and comparing opinions on the web. *WWW '05: Proc. of the 14th int. conf. on World Wide Web*, New York, NY, USA, 2005, ACM Press, sivut 342–351.

[Ni07] Ni, X., Xue, G.-R., Ling, X., Yu, Y. ja Yang, Q., Exploring in the weblog space by detecting informative and affective articles. *WWW '07: Proc. of the 16th int. conf. on World Wide Web*, New York, NY, USA, 2007, ACM Press, sivut 281–290.

[Son04] Song, R., Liu, H., Wen, J.-R. ja Ma, W.-Y., Learning block importance models for web pages. *WWW '04: Proc. of the 13th int. conf. on World Wide Web*, New York, NY, USA, 2004, ACM Press, sivut 203–211.

[Tur02] Turney, P. D., Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. *ACL '02: Proc. of the 40th Annual Meeting on Association for Computational Linguistics*, Morristown, NJ, USA, 2002, Association for Computational Linguistics, sivut 417–424.

[YuH03] Yu, H. ja Hatzivassiloglou, V., Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. *Proc. of the 2003 conf. on Empirical methods in natural language processing*, Morristown, NJ, USA, 2003, Association for Computational Linguistics, sivut 129–136.

[YiN05] Yi, J. ja Niblack, W., Sentiment mining in webfountain. *ICDE '05: Proc. of the 21st Int. Conf. on Data Engineering (ICDE'05)*, Washington, DC, USA, 2005, IEEE Computer Society, sivut 1073–1083.

[ZJZ06] Zhuang, L., Jing, F. ja Zhu, X.-Y., Movie review mining and summarization. *CIKM '06: Proc. of the 15th ACM int. conf. on Information and knowledge management*, New York, NY, USA, 2006, ACM Press, sivut 43–50.

[ZhV06] Zhang, Z. ja Varadarajan, B., Utility scoring of product reviews. *CIKM '06: Proc. of the 15th ACM int. conf. on Information and knowledge management*, New York, NY, USA, 2006, ACM Press, sivut 51–57.