WordNet-sanatietokannan hyödyntäminen luonnollisen kielen sovelluksissa

Anne Varis

Helsinki 12.1.2004

Pro gradu -tutkielma

HELSINGIN YLIOPISTO
Tietojenkäsittelytieteen laitos

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta/Osasto — Fakultet/Sektion — Faculty	Laitos — Institution — Department
Matemaattis-luonnontieteellinen	Tietojenkäsittelytieteen laitos

Tekijä — Författare — Author
Anne Varis

Työn nimi — Arbetets titel — Title
WordNet-sanatietokannan hyödyntäminen luonnollisen kielen sovelluksissa

Oppiaine — Läroämne — Subject
Tietojenkäsittelytiede

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Pro gradu -tutkielma	12.1.2004	88 sivua

Tiivistelmä — Referat — Abstract

Tutkielmassa käsitellään WordNet -sanatietokannan hyödyntämistä luonnollisen kielen sovelluksissa. Näitä sovelluksia ovat disambiguaatio, verbien aspektuaalinen yksiselitteistäminen, tiedon haku, dokumenttien luokittelu ja leksikaalisten ketjujen muodostaminen.

Tutkimuksen perusteella sanatietokannan käytöstä näyttäisi olevan eniten hyötyä dokumenttien luokittelussa ja leksikaalisten ketjujen muodostamisessa. Hyödyllisyys riippuu osittain sovelluksesta ja käsiteltävien tekstien aihepiiristä.

ACM Computing Classification System (CCS): H.3.1, H.3.3, I.1.2

Avainsanat — Nyckelord — Keywords
WordNet, yksiselitteistäminen, tiedon haku, dokumenttien luokittelu, leksikaaliset ketjut

Säilytyspaikka — Förvaringsställe — Where deposited
Tietojenkäsittelytieteen laitoksen kirjasto, sarjanumero C-2004-

Muita tietoja — övriga uppgifter — Additional information

Sisältö

1 Johdanto

Tässä tutkielmassa tarkastellaan WordNet-sanatietokannan hyödyntämistä luonnollisen kielen sovelluksissa (Natural Language Applications). Näistä sovelluksista käytetään tässä tutkielmassa useimmiten lyhennettä NLP-sovellukset. Näitä sovellusalueita ovat esimerkiksi disambiguaatio, tiedonhaku, dokumenttien luokittelu ja leksikaalisten ketjujen muodostaminen.

WordNet on englannin kielen sanojen semanttisten suhteiden tietokanta. WordNetin kehittämisessä päämääränä on ollut luoda tietokanta, joka paitsi kuvaisi semanttisten suhteiden järjestäytymistä ihmisen mielessä, olisi käyttökelpoinen tietokoneohjelmistoille. WordNetia kuvataan tarkemmin luvussa 2.

Disambiguaatio tarkoittaa sitä, että moniselitteiselle sanalle valitaan merkitys sanan kontekstin perusteella. Disambiguaatiosta puhutaan tässä tutkielmassa useimmiten yksiselitteistämisenä. Disambiguaatiomenetelmistä käytetään tässä tutkielmassa lyhennettä WSD- (Word Sense Disambiguation) menetelmät.

Esimerkkinä sanan merkityksen valinnasta sen kontekstin perusteella on sana *plant*, jolla on merkitykset *plant* (kasvi) ja *plant* (tehdas) [Yar95]. Se voi esiintyä esimerkiksi seuraavanlaisissa virkkeissä:

> Company said the *plant* is still operating.
> ... divide life into *plant* and animal kingdom.

Sanan merkitys voidaan tässä tapauksessa päätellä lähellä olevien sanojen perusteella. Merkityksen (tehdas) tapauksessa tälläinen vihjesana voisi olla esimerkiksi *company*, merkityksen (kasvi) tapauksessa vihjesana voisi olla *animal*.

Mihalcea ja Moldovan [MM98] jakavat WSD-menetelmät kolmeen päätyyppiin:

1) Elektroniseen sanakirjaan perustuvat menetelmät.

2) Ohjattuun koneoppimiseen perustuvat menetelmät, joissa käytetään manuaalisesti luokiteltua opetusaineistoa.

3) Ohjaamattomaan koneoppimiseen perustuvat menetelmät, joissa käytetään manuaalisesti luokittelematonta opetusaineistoa.

Jotkut WSD-menetelmät perustuvat osittain koneoppimiseen ja osittain elektronisen sanakirjan hyödyntämiseen [LCM98, MM98, WOB98].

Koneoppimiseen perustuvilla WSD-menetelmillä on raportoitu voitavan yksiselitteistää oikein jopa yli 90 prosenttia testiaineistossa esiintyvistä moniselitteisistä sanoista [Yar95]. Elektroniseen sanakirjaan perustuvilla menetelmillä ei ole raportoitu saadun yhtä hyviä tuloksia. Niillä on kuitenkin käyttöä joissakin interaktiivisissa NLP-sovelluksissa, joissa ei ole käytettävissä manuaalisesti luokiteltua oppimiskorpusta, ja/tai koneoppimisprosessiin ei ole aikaa.

Verbien aspektuaalinen luokittelu tarkoittaa sitä, että verbi liitetään oikeaan aspektiluokkaan. Yleisesti oletetaan, että on olemassa kolme aspektiluokkaa: tila (state), toiminta (activity) ja tapahtuma (event) [Ven67, Pus95]. Tapahtuma-luokka voidaan edelleen jakaa suorituksiin (accomplishment) ja saavutuksiin (achievement). Aspektuaalista luokittelua tarvitaan malleissa, jotka arvioivat temporaalisia rajoituksia lauseiden välillä [MS88, Sie98].

Tiedonhaussa etsitään tyypillisesti dokumentteja, joissa esiintyy kyselyssä esiintyviä sanoja. Elektronista sanakirjaa hyödyntämällä voidaan suorittaa hakuja sanamerkityksillä leksikaalisten sanojen sijaan. Toinen tapa hyödyntää elektronista sanakirjaa tiedon haussa on kyselyjen rikastaminen. Dokumenttien luokittelussa elektronisesta sanakirjasta saatavaa informaatiota voidaan käyttää koneoppimisinformaation täydennyksessä. Leksikaalisten ketjujen muodostamisessa elektronista sanakirjaa hyödynnetään leksikaalisena resurssina, joka järjestää sanat niiden merkityksen mukaan.

Tutkielman päämääränä on esitellä elektronista sanakirjaa, tässä tapauksessa Word-

Netia hyödyntäviä NLP-sovelluksia. Tutkielmassa tarkastellaan leksikaalisen tietokannan, tässä tapauksessa WordNetin hyödyllisyyttä erilaisten NLP-sovellusten tapauksissa ja sitä, kuinka WordNetin ominaisuudet vaikuttavat sen käytettävyyteen erilaisissa sovelluksissa. Nämä ominaisuudet voivat liittyä esimerkiksi WordNetin rakenteeseen tai kattavuuteen.

Luvussa 2 kuvataan yleisesti WordNet-tietokantaa. Luvussa 3 tarkastellaan WordNetia hyödyntäviä WSD-menetelmiä, joiden avulla yksiselitteistetään moniselitteisiä sanoja, yleensä substantiiveja. Verbien aspektuaalista yksiselitteistämistä käsitellään erikseen luvussa 4. Luvun 5 aiheena on leksikaalisten ketjujen muodostaminen WordNetin avulla. Luvussa 6 käsitellään WordNetin hyödyntämistä tiedonhakusovelluksissa. Luvussa 7 tarkastellaan eri tapoja hyödyntää WordNetia dokumenttien luokittelussa.

2 WordNet

2.1 Yleistä

WordNet on etenkin tutkimusmaailmassa paljon käytetty englannin kielen sanojen semanttisten suhteiden tietokanta. Alkuperäinen WordNet on kehitetty Yhdysvalloissa Princetonin yliopiston kognitiotieteen laitoksella. Työ aloitettiin 1985. WordNet on vapaasti saatavissa tutkimuskäyttöön ja se on myös käytettävissä www-lomakkeiden kautta Princetonin yliopiston kognitiotieteen laitoksen www-sivuilla: *http://www.cogsci.princeton.edu/~wn/* . Seuraavaksi esitetään WordNetin rakenne ja perusperiaatteet yleisesti.

WordNet koostuu kolmesta erillisestä tietokannasta. Yhdessä ovat substantiivit, toisessa verbit ja kolmannessa adjektiivit ja adverbit. Tietokannassa ovat siis edustettuina kaikki avoimet sanaluokat, eli sanaluokat, joihin syntyy helposti uusia sanoja. Suljettuihin sanaluokkiin kuuluvia sanoja ei WordNetissa ole lainkaan. Sulkeisia sanaluokkia ovat pronominit, prepositiot, postpositiot, artikkelit ja konjunktiot. Jos jokin perusmuoto kuuluu useampaan sanaluokkaan, se on erikseen talletettu kaikkiin niihin tietokantoihin, joihin se sanaluokkansa puolesta kuuluu.

WordNetin kehittämisessä oli päämääränä luoda sellainen sanojen semanttisten suhteiden tietokanta, joka paitsi kuvaisi semanttisten suhteiden järjestäytymistä ihmisen mielessä, olisi käyttökelpoinen tietokoneohjelmistoille. Aikaisemmin olemassa olevat sanakirjat ja sanastot oli suunniteltu ihmisen luettaviksi, eivätkä ne sellaisenaan soveltuneet tietokoneohjelmien käytettäviksi.

Synonymia on semanttinen perussuhde WordNetissa. WordNetissa käytetty synonymian käsite ei tarkoita sitä, että sanat olisivat vaihdettavissa keskenään kaikissa konteksteissa. Riittää, että ne ovat vaihdettavissa keskenään joissakin konteksteissa. [Mil98a] WordNetin yhteydessä ei oikeastaan voida puhua sanoista. Sen takia

tässä tutkielmassa käytetään termiä perusmuoto WordNetiin talletetuista yksiköistä [Aun02]. WordNetin perusmuoto voi koostua useammasta kuin yhdestä sanasta, esimerkiksi *atom bomb*, *interpretative dancing* tai *break down*. WordNetin perusmuoto ei ole sama kuin lekseemi (sanan leksikkomuoto), sillä homonyymit ovat WordNetissa yksi ja sama perusmuoto, jos ne kuuluvat samaan tietokantaan, esimerkiksi *bass* (basso) ja *bass* (ahven). WordNet ei myöskään tee eroa homonymian ja polysemian välillä. Kaksi sanamerkitystä ovat homonyymisia, jos ne ovat ilmiasultaan samoja, mutta niiden merkitykset eivät ole semanttisesa suhteessa toisiinsa, kuten edellisessä *bass* esimerkissä. Kaksi sanamerkitystä ovat polyseemisia, jos ne ovat ilmiasultaan samoja ja niiden merkitykset ovat jossakin semanttisessa suhteessa toisiinsa, esimerkiksi pullon suu ja joen suu.

WordNetin semanttiset suhteet vallitsevat perusmuotojen välillä, perusmuotojen yksittäisten merkitysten välillä tai synonyymijoukkojen välillä. Kaikki WordNetissa esiintyvät semanttiset suhteet on esitetty taulukossa 1. Synonyymijoukko kuvaa käsitettä, jonka oletetaan olevan ihmisen leksikaalisessa muistissa. Taulukossa 2 on esitetty perusmuodolle *dog* substantiivitietokannasta löytyvistä kuudesta merkityksestä kolme ensimmäistä. Taulukossa 2 esiintyvät synonyymijoukot ovat { *dog*, *domestic dog*, *canis familiaris* } ja {*fromp dog*}. Perusmuodon eri merkitykset WordNetissa on järjestetty yleisyyden perusteella siten, että useimmin esiintyvä on ensimmäisenä ja harvimmin esiintyvä viimeisenä.

WordNet sisältää yli 166 000 yksittäistä perusmuodon merkitystä, yli 118 000 perusmuotoa ja yli 90 000 merkitystä eli synonyymijoukkoa. Noin 17 prosentilla WordNetin perusmuodoista on useampi kuin yksi merkitys, ja noin 40 prosentilla yksi tai useampi synonyymi. WordNetissa on yhteensä noin 116 000 semanttista suhdetta osoittavaa viittausta perusmuotojen, perusmuotojen yksittäisten merkitysten ja synonyymijoukkojen välillä.

Suhde	Tietokanta	Esimerkki
Synonymia	N, V, Adj/Adv	sad, unhappy
Antonymia	Adj/Adv, (N, V)	sad, glad (sell, buy)
Hyponymia hyperonyymi hyponyymi	N	silver maple is a kind of *maple* *silver maple* is a kind of maple
Meronymia holonyymi meronyymi	N	ship is a member of a *fleet* *ship* is a member of a fleet
Troponymia troponyymi	V	to whisper is a particular way to *speak*
Implikaatio (entailment)	V	to win entails *competing*

Taulukko 1: *WordNetissa esiintyvät semanttiset suhteet*

1. *dog*, domestic dog, Canis familiaris - (a member of genus Canis (probably descented from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds; "the dog barked all night");
2. fromp, *dog* - (a dull unattractive unpleasant girl or woman; "she got a reputation as a fromp"; "she is a real dog")
3. dog - (informal term for a man: "you lucky dog")

Taulukko 2: *Sanan dog kolme ensimmäistä substantiivitietokannasta löytyvää merkitystä*

2.2 Substantiivitietokanta

Kaikki WordNetissa esiintyvät substantiivit on järjestetty hierarkiaksi hyponymian mukaan. Substantiivin hyponyymi on sen aliluokka ja suhde hyponyymin ja hyperonyymin välillä on *kind-of* suhde. Esimerkiksi *maple* (vaahtera) on sanan *tree* (puu) hyponyymi, eli *maple* on eräänlainen puu. Hyperonymia ja hyponymia ovat transitiivisia semanttisia suhteita synonyymijoukkojen välillä. Puun juurisolmulla on 25 lapsisolmua, jotka on lueteltu taulukossa 3. Vaikka hierarkia puun juuresta yksittäiseen lehtisolmuun voi olla teoriassa kuinka monikerroksinen tahansa, se on vain harvoin yli 10 tasoa.

{act, action, activity}	{animal, fauna}	{artifact}
{attribute, property}	{body, corpus}	{cognition, knowledge}
{communication}	{event, happening}	{feeling, emotion}
{food}	{group, collection}	{location, place}
{motive}	{natural object}	{natural phonemenon}
{person, human being}	{plant, flora}	{possession}
{process}	{quantity, amount}	{relation}
{shape}	{state, condition}	{substance}
{time}		

Taulukko 3: *WordNetin substantiivihierarkian juurisolmun {entity} lapsisolmut*

Muita substantiivitietokannassa esiintyviä suhteita ovat *meronymia* (osa-kokonaisuus) ja *antonymia* (semanttinen oppositio) [MBF+93]. Meronyymit ovat piirteitä, joita hyponyymit voivat periä. Tämän takia meronymia ja hyponymia ovat kompleksisessa suhteessa keskenään. Esimerkiksi, koska *bak* (nokka) ja *wing* (siipi) ovat sanan *bird* (lintu) meronyymeja, niiden täytyy olla myös sanan *canary* meronyymeja. Semanttinen oppositio ei ole perustavaa laatua oleva suhde substantiivien välillä, mutta sillä on oma esitystapansa WordNetissa. Esimerkiksi sanat *man* ja *woman* ovat semanttisessa

oppositiossa keskenään.

2.3 Verbitietokanta

Verbit on WordNetissa järjestetty pääosin verbiluokkien, merkityksen ja leksikaalisten implikaatiosuhteiden (entailment relations) mukaan. Verbien väliset suhteet ja verbien järjestys WordNetissa ovat monimutkaisemmat kuin substantiivien tai adjektiivien ja adverbien keskinäiset suhteet ja järjestys. Tyypilliset verbihierarkiat ovat suhteellisen matalia. Niiden syvyys ylittää harvoin neljä tasoa.

Verbiluokkien perusteella verbit on jaettu ensin kahteen luokkaan: tapahtumiin ja toimintaan (events and actions), sekä tiloihin (states). Sen jälkeen tapahtuma- ja toimintaluokkaan kuuluvat verbit on jaettu edelleen neljääntoista eri luokkaan. Niitä vastaavat verbihierarkiapuun solmut on esitetty taulukossa 4.

bodily care and functions	change	cognition
communication	competition	consumption
contact	creation	emotion
motion	perception	possession
social interaction	weather	

Taulukko 4: *WordNetin verbihierarkian tapahtumasolmun ja toimintasolmun lapsisolmut*

Tilaluokkaan luokitellut verbit eivät muodosta mitään yhtenäistä semanttista aluetta, vaan niitä yhdistää se, että ne eivät kuulu mihinkään tapahtumaluokan ja toimintaluokan 14 aliluokasta. Esimerkkejä tilaluokan verbeistä ovat *suffice* (riittää), *belong* (kuulua jollekin) ja *resemble* (muistuttaa jotakin). Merkityksensä perusteella verbit on järjestetty hierarkkisesti siten, että hierarkian ylimmällä tasolla on juuriverbejä eli verbejä, joiden merkitys on osa alemmalla tasolla olevien verbien merkityksestä.

Juuriverbejä ovat esimerkiksi *move*, *go* ja *change*, koska ne kuvaavat peruskäsitteitä. Esimerkiksi *walk* ja *run* sisältävät juuriverbien *move* ja *go* merkitykset.

Kolmas tapa järjestää verbejä ja kuvata niiden välisiä suhteita on leksikaalinen implikaatio, jolla tarkoitetaan sitä, että verbin V1 kuvaamasta toiminnasta seuraa verbin V2 kuvaama toiminta. Leksikaalinen implikaatio on sama kuin propositiologiikan implikaatio P -> Q. Esimerkiksi *kuorsaaminen* implikoi *nukkumista*. Jos henkilö kuorsaa, siitä voidaan päätellä, että hän nukkuu. Leksikaalinen implikaatiosuhde on edelleen jaettu neljään osaan: troponymiaan, oletukseen ajassa taaksepäin (backward presupposition) , syy-seuraus-suhteeseen (cause) ja muihin samanaikaisuutta sisältäviin implikaatiosuhteisiin. Seuraavaksi kuvataan tarkemmin kutakin implikaatiosuhdetta. Verbien V1 ja V2 välinen implikaatiosuhde on:

Troponymia, jos V1 ilmaisee tapaa tehdä V2. Esim. V1 = *limp* (nilkuttaa), V2 = *walk*

Oletus ajassa taaksepäin, jos V1 ei voi tapahtua, ilman että sitä ennen on tapahtunut V2. Esim. V1 = *go back*, V2 = left.

Syy-seuraus, jos V1 on kausatiivinen (causative) ja V2 on resultatiivinen (resultative). Esim. V1 = teach, V2 = learn.

Muu samanaikaisuutta sisältävä implikaatiosuhde, jos V1 ja V2 tapahtuvat ainakin osittain samanaikaisesti, eikä niiden välinen implikaatiosuhde ole mikään edellä mainituista suhteista. Esim. V1 = *buy* (ostaa), V2 = *pay* (maksaa).

Jotta verbien tärkeimmät syntaktiset ominaisuudet tulisivat kuvatuiksi, WordNet sisältää kullekin verbisynonyymijoukolle yhden tai useampia virkekehyksiä (sentence frames), jotka spesifioivat synonyymijoukon verbien tärkeimmät alikategorisaatiopiirteet (subcategorization features) osoittamalla, millaisissa konteksteissa ne voivat esiintyä. Esimerkkinä ovat sanan *run* merkityksen 1 virkekehykset:

run – (move fast by using one's feet, with one foot off the ground at any given time; "Don't run–you'll be out of breath"; "The children ran to the store")

2.4 Adjektiivi- ja adverbitietokanta

WordNetissa adjektiivit on jaettu kahteen pääluokkaan, kuvaileviin ja relationaalisiin adjektiiveihin [LCM98]. Kuvailevat adjektiivit on organisoitu binääristen oppositioiden (antonymioiden) ja merkitysten samanlaisuuden (synonymioiden) mukaan. Niillä kuvailevilla adjektiiveilla, joilla ei ole suoria antonymioita, voi olla epäsuoria antonymioita, johtuen niiden semanttisesta samankaltaisuudesta adjektiivien kanssa, joilla on suoria antonymioita. Esimerkki adjektiivista, jolla ei ole suoria antonymioita on *sultry* (polttava). Sen merkitys on samankaltainen, kuin adjektiivin *hot*, jolla on suora antonyymi *cold*. Adjektiivin *cold* on *sultry*:n epäsuora antonyymi. Kuvailevat adjektiivit muuttavat pääsanana toimivan substantiivin merkitystä.

Relationaaliset adjektiivit, jotka ovat substantiiveista johdettuja adjektiiveja, on WordNetissa yhdistetty kyseisiin subtantiiveihin. Esimerkiksi perusmuoto *dental* on suhteessa sanaan *tooth*. Värejä ilmaisevat sanat on WordNetissa kuvattu eri tavalla kuin muut adjektiivit, sillä englannin kielen väriadjektiivit ovat monella tavoin poikkeuksellisia. Ne voivat toimia joko substantiiveina tai adjektiiveina, mutta ovat kuitenkin adjektiiveja. Niitä voidaan käyttää kuten muitakin kuvailevia adjektiiveja. Muihin kuvaileviin adjektiiveihin sopiva suoran ja epäsuoran antonymian malli ei kuitenkaan sovi väriadjektiiveihin.

Useimmat englannin kielen adverbit on johdettu adjektiiveista suffiksien avulla, esimerkiksi *add -> addly*. Tälläiset adverbit on WordNetissa liitetty niihin adjektiiveihin, joista ne on johdettu [Mil98b].

3 WordNetia hyödyntävät substantiivien yksiselitteistämismenetelmät

Tässä luvussa tarkastellaan moniselitteisten sanojen, yleensä substantiivien, yksiselitteistämiseen käytettyjä WSD-menetelmiä. Luvussa 3.1 kuvataan lyhyesti semanttisen etäisyyden ja semanttisen tiheyden käsitteitä. Luvussa 3.2 tarkastellaan Aqirren ja Rigaun [AR95, AR96] esittelemää WSD-menetelmää, joka käyttää informaation lähteenä pelkästään WordNet-tietokantaa. Luvussa 3.3 käsitellään lyhyesti WordNetia hyödyntäviä tilastollisia yksiselitteistämismenetelmiä. Luvun 3.4 aiheena ovat substantiiviryhmien yksiselitteistämismenetelmät.

3.1 Semanttinen etäisyys ja semanttinen tiheys

Substantiivien yksiselitteistämisessä käytettävät, WordNetin IS-A- hierarkiaa hyödyntävät WSD-menetelmät hyödyntävät tyypillisesti yksiselitteistämisessä *semanttista* (tai käsitteellistä) *etäisyyttä* ja *semanttista tiheyttä*. Rada [RMBB89] määrittelee kahden käsitteen välisen käsitteellisen etäisyyden lyhyimmän polun pituudeksi, joka yhdistää käsitteitä hierarkkisessa semanttisessa verkossa. Aqirren ja Rigaun [AR95, AR96] mukaan käsitteellisen etäisyyden mittaamiseen käytetyissä kaavoissa on huomioitava seuraavat asiat:

1) Lyhyimmän polun pituus, joka yhdistää kyseisiä käsitteitä.

2) Hierarkian syvyys. Hierarkian syvemmässä osassa olevat käsitteet tulee katsoa toisilleen läheisemmiksi.

3) Käsitteellinen tiheys hierarkiassa. Hierarkian tiheässä osassa olevat käsitteet ovat käsitteellisesti lähempänä toisiaan kuin harvassa osassa olevat.

4) Mittaamisen tulisi olla riippumatonta mitattavien käsitteiden määrästä.

Mihalcea ja Moldovan määrittelevät semanttisen tiheyden niiden yhteisten sanojen määräksi, jotka ovat kahden tai useamman sanan semanttisen etäisyyden päässä kohdesanoista [MM98]. Mitä läheisempi semanttinen suhde kahden sanan välillä on, sitä suurempi semanttinen tiheys niiden välillä on.

3.2 Pelkästään WordNetia hyödyntävä yksiselitteistämismenetelmä

Tässä aliluvussa tarkastellaan Aqirren ja Rigaun [AR95, AR96] esittelemää WSD-menetelmää. Tämä menetelmä hyödyntää yksinomaan WordNetia. Sitä on testattu yleisessä käytössä olevalla tekstiaineistolla SemCor. SemCor on manuaalisesti, WordNet- merkitysten mukaan luokiteltu osa Brownin korpuksesta. Brownin korpus on Brownin yliopistossa kehitetty miljoonan sanan korpus, joka sisältää amerikan englannin tekstejä. Korpus koostuu viidestäsadasta, viidestätoistä eri tekstikategoriasta kootusta tekstistä.

Menetelmä perustuu WordNetin substantiivitaksonomiaan ja käsitteelliseen etäisyyteen käsitteiden välillä. Tätä etäisyyttä kuvaa tähän tarkoitukseen kehitetty käsitteellisen tiheyden kaava. Menetelmä on täysin automaattinen, eikä se vaadi leksikaalisten yksiköiden koodaamista käsin, oppimisaineiston manuaalista luokittelua eikä minkäänlaista koneoppimisprosessia. Käytetyssä käsitteellisen tiheyden laskentakaavassa c on alihierarkian huipulla oleva käsite ja m on yksiselitteistettävän sanan merkitysten määrä:

$$\mathrm{CD}(c,m) = \frac{\sum_{i=0}^{m-1} nhyp^{i^{0.20}}}{descendants_c}$$

nhyp on solmujen hyponyymien määrän keskiarvo ja $descendants_c$ on c:n jälkeläisten määrä.

WSD-järjestelmän täytyy tietää, kuinka sanat on klusteroitu semanttisiin luokkiin, ja kuinka semanttiset luokat on organisoitu hierarkkisesti. Tämä tieto saadaan WordNetista. Syötteeksi annetaan tekstikatkelma. Järjestelmä yrittää ratkaista substantiivien leksikaalisen moniselitteisyyden löytämällä jostain substantiivijoukosta sellaisen merkitysten yhdistelmän, joka maksimoi merkitysten välisen totaalisen käsitteellisen tiheyden. WordNetin kattavuus tässä tapauksessa oli noin 89 prosenttia, eli noin 89 prosenttia käsitellyistä substantiiveista löytyi WordNetista.

3.2.1 Algoritmi

Ohjelma siirtää ikkunaa substantiivi kerrallaan dokumentin alusta loppuun. Jokaisella askeleella yksiselitteistetään ikkunan keskellä oleva substantiivi ja huomioidaan muut ikkunassa olevat substantiivit kontekstina. Ikkunan koko on viisi.

Yksiselitteistämisalgoritmi etenee seuraavasti. Ensiksi algoritmi esittää ikkunassa olevien substantiivien WordNetin mukaiset merkitykset ja hyperonyymit. Seuraavaksi ohjelma laskee kunkin käsitteen käsitteellisen tiheyden sen alihierarkiassa olevien merkitysten perusteella. Algoritmi valitsee käsitteen c, jonka tiheys on suurin, ja valitsee sen alla olevat merkitykset vastaavien sanojen korrekteiksi merkityksiksi. Jos ikkunassa olevalla sanalla on vain yksi merkitys c:n alla, se on jo yksiselitteistetty. Jos sillä ei ole yhtään merkitystä c:n alla, se on yhä moniselitteinen. Jos sillä on useampi kuin yksi merkitys c:n alla, muut kuin c:n alla olevat merkitykset voidaan eliminoida, mutta sanaa ei ole vielä täydellisesti yksiselitteistetty.

Algoritmi etenee seuraavaksi muihin ikkunassa oleviin substantiiveihin laskeakseen niille käsitteellisen tiheyden ja yksiselitteistääkseen ne. Kun yksiselitteistämistä ei ole enää mahdollista jatkaa, sanan mahdolliset jäljellä olevat merkitykset käsitellään, ja tulos esitetään. Prosessin havainnollistamiseksi tarkastellaan seuraavaa SemCorista peräisin olevaa tekstikatkelmaa:

> The *jury(2)* praised the *administration(2)* and *operation(8)* of the Atlanta *Police Department(1)*, the Fulton-Tax-Commissioner's-Office, the Bellwood and Alphanetta *prison-farms(1)*, Grady-Hospital and the Fulton-Health-Department

Sanat *Police Department, jury, operation* ja *administration* ovat WordNetissa esitettyjä substantiiveja. Niiden merkitysten määrät on annettu suluissa. Tässä esimerkissä yksiselitteistettävä sana on *operation* ja ikkunan koko on viisi.

Askel 1

Ikkunassa olevien sanojen mahdolliset merkitykset etsitään WordNetista. Kuvassa 1 on esitetty osittainen WordNet-hierarkia esimerkkivirkkeelle. Koska *prison-farm* kuuluu eri hierarkiaan, sitä ei käsitellä tässä yhteydessä. Yksiselitteistettäville substantiiveille on annettu merkitysnumerot. Yksiselitteisillä substantiiveilla on merkitysnumero 0.

Askel 2

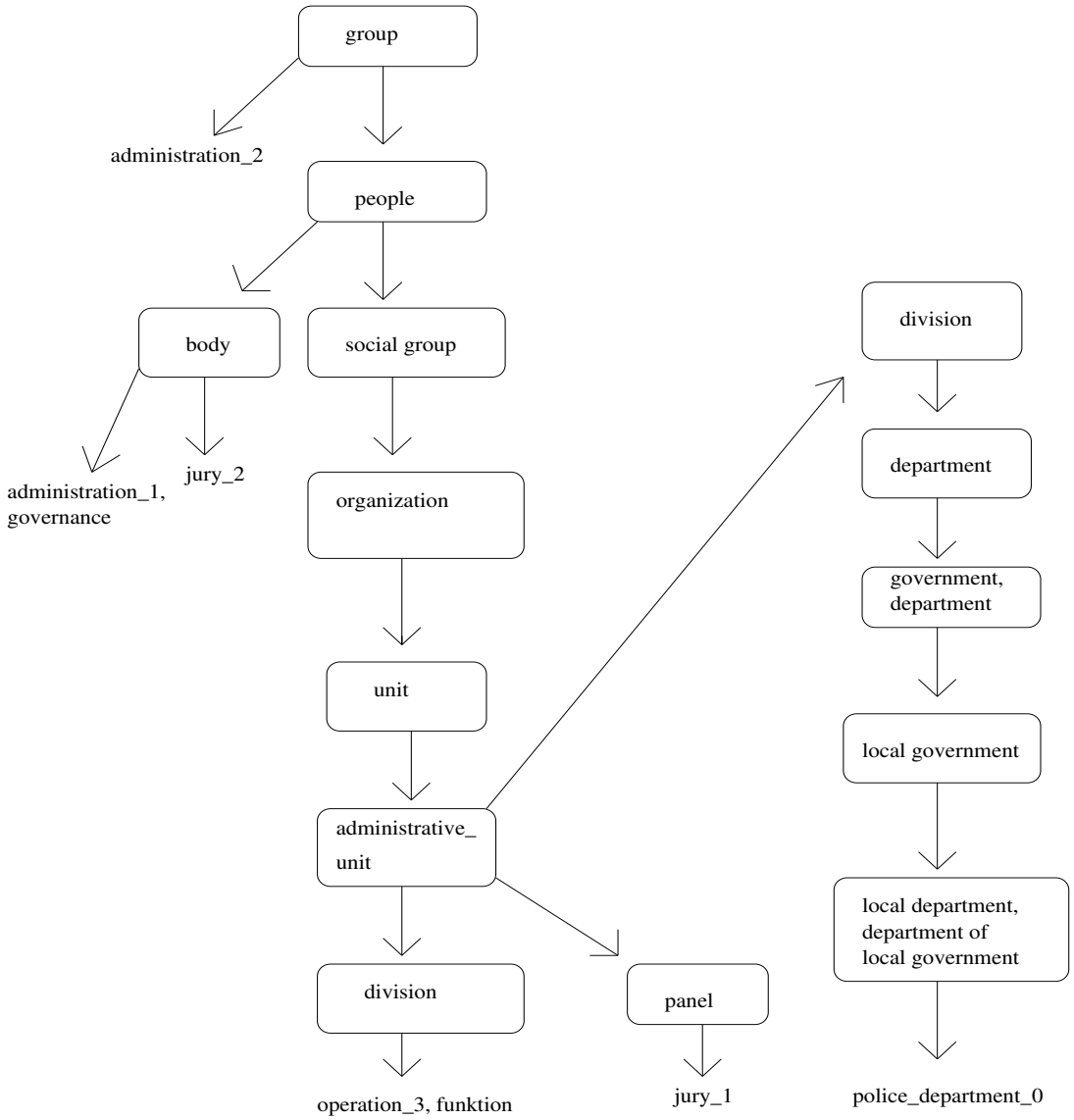
Merkitysten käsitteelliset tiheydet lasketaan. Sanalla <administrative-unit> on kolme mahdollista merkitystä, ja sen alihierarkian koko on 96. Näin ollen sen käsitteelliseksi tiheydeksi tulee 0.256. Sana <body>, jolla on kaksi merkitystä, ja jonka alihierarkian koko on 86, saa käsitteelliseksi tiheydeksi 0.062.

Askel 3

<administrative-unit> valitaan suurimman käsitteellisen tiheyden omaavaksi käsitteeksi.

Askel 4

Suurimman käsitteellisen tiheyden omaavan käsitteen alla olevat merkitykset valitaan käsiteltävien sanojen merkityksiksi. $operation_3$, $police\_department_0$ ja $jury_1$ ovat merkitykset, jotka on valittu sanoille *operation*, *Police Department* ja *jury* . Kaikki



Kuva 1: Esimerkkitekstin sanojen osittainen WordNet-hierarkia

muut käsitteen <administrative-unit> alla olevat käsitteet on merkitty niin, että niitä ei enää valita. Näiden sanojen muut merkitykset esim. $jury_2$ poistetaan. Seuraavalla kierroksella sanalla <body> on alapuolellaan vain yksi yksiselitteistettävä sana, ja sen käsitteellinen tiheys on täten huomattavasti alhaisempi. Tässä vaiheessa algoritmi päättelee, että yksiselitteistäminen ei ole enää mahdollista ja poistuu silmukasta.

Askel 5

Algoritmi on yksiselitteistänyt käsitteet $operation_3$, $police\_department_0$, $jury_1$ ja $prison\_farm_0$, mutta sana *administration* on yhä moniselitteinen. Algoritmin tulos on se, että sanan *operation* merkitys tässä kontekstissa (eli tässä ikkunassa) on $operation_3$. Ikkuna siirtyy oikealle, ja algoritmi yrittää yksiselitteistää sanan *Police Department*, huomioiden kontekstina sanat *administration*, *operation*, *prison farms* ja seuraavan virkkeen ensimmäisen substantiivin.

Algoritmi voi joko onnistua yksiselitteistämään sanan, epäonnistua siinä tai palauttaa useita mahdollisia merkityksiä.

Tehdyissä testeissä algoritmi yksiselitteisti oikein 71.2 prosenttia tekstissä esiintyvistä polyseemisista substantiiveista..

Aqirren ja Rigaun mukaan seuraavat tekijät voisivat parantaa algoritmin suorituskykyä:

1) Koherenttien tekstiyksiköiden työstäminen kerrallaan. SemCorissa ei ole mitään diskurssin rakennetta koskevaa informaatiota, lukuun ottamatta virkkeiden loppuja. Jos syötteenä olisi koherentteja diskurssin osia, algoritmin suorituskyky ja tehokkuus todennäköisesti paranisivat. Suorityskyky voisi parantua, jos keskenään eri aihepiireihin kuuluvia virkkeitä ei käsiteltäisi yhdessä yksiselitteistämisikkunassa. Tehokkuus voisi olla parempi, jos algoritmi työstäisi koko tekstiyksikön kerralla, sen sijaan, että työstettäisiin sana kerrallaan.

2) Semanttisen datan laajentaminen. WordNet tarjoaa synonymia-, hyponymia- ja

meronymia-suhteet substantiiveille. WordNetista puuttuvat esimerkiksi sanaluokkien väliset semanttiset suhteet. Paitsi että tälläiset suhteet mahdollistaisivat verbien, adjektiivien ja adverbien yksiselitteistämisen, ne voisivat kuvata paremmin suhteita merkitysten välillä ja tarjota paremman perustan yksiselitteistämiselle.

3.3 Tilastolliset yksiselitteistämismenetelmät

Sekä WordNetia että tilastoja hyödyntävissä WSD-menetelmissä haetaan tyypillisesti sanan mahdolliset merkitykset WordNetista ja kootaan sitten oppimiskorpuksesta tilastoja eri merkitysten todennäköisyyksistä [LCM98, MM98, WOB98].

Mihalcean ja Moldovanin [MM98] esittelemässä menetelmässä sanalle etsitään mahdolliset merkitykset WordNetista ja ne asetetaan todennäköisyysjärjestykseen oppimiskorpuksesta koottujen tilastojen perusteella. Leacockin et. al. [LCM98] esittelemässä menetelmässä käytetään WordNetista saatua leksikaalista informaatiota oppimisesimerkkien eristämiseen luokittelemattomasta opetusaineistosta. Wieben et.al. [WOB98] esittelemässä lähestymistavassa eristetään WordNetin IS-A-hierarkiasta virkkeessä olevien kohdesanojen synonyymijoukot ja ne synonyymijoukot, jotka ovat saavutettavissa niistä. Tästä informaatiosta muodostetaan Bayes-verkko. Tämän jälkeen kehitetään empiirinen luokittelija jokaiselle moniselitteiselle sanalle. Jokainen luokittelija määrittelee todennäköisyysjakauman, joka kuvaa kunkin kohdesanan merkityksen todennäköisyyttä, ottaen huomioon kontekstuaaliset piirteet.

Toinen Mihalcean ja Moldovanin [RM00] esittämä WordNetia ja tilastoja hyödyntävä WSD-menetelmä määrittelee iterativiisen algoritmin avulla, mitkä kohdetekstin substantiiveista ja verbeistä voidaan yksiselitteistää tarkasti. Substantiiveista ja verbeistä yksiselitteistetään noin 55 prosenttia, mutta se voidaan tehdä hyvin tarkasti (noin 92 prosentin tarkkuudella).

3.4 Substantiiviryhmien yksiselitteistämismenetelmät

Tässä aliluvussa tarkastellaan lyhyesti Sussnan [Sus93] ja Resnikin [Res95] esittelemiä substantiiviryhmien yksiselitteistämismenetelmiä. Substantiiviryhmät muodostetaan joko substantiiveista, jotka ovat lähellä toisiaan tekstissä [Sus93], käyttämällä klusterointialgoritmia tai manuaalisesti [Res95]. Ideana on, että ryhmän muiden substantiivien merkitykset auttavat kunkin yksittäisen substantiivin yksiselitteistämisessä.

Sussna [Sus93] esittää substantiiviryhmien yksiselitteistämisalgoritmin tiedonhakua varten. Substantiiviryhmä muodostetaan substantiiveista, jotka ovat lähellä toisiaan syntaktisesti jäsennetyssä tekstissä. Menetelmässä sanojen yhteenkuuluvuus kuvataan semanttisen verkon termein. Semanttisen samankaltaisuuden tai semanttisen etäisyyden arviointi on merkityksen valinnan perustana. IS-A-hierarkian lisäksi käytetään muitakin WordNet-suhteita, kuten PART-OF. Semanttinen samankaltaisuus tai etäisyys perustuu polun pituuteen.

Syötetekstin termit, joilla on monia merkityksiä yksiselitteistetään etsimällä joukosta läheisiä termejä merkitysyhdistelmä, joka minimoi totaalisen etäisyyden merkitysten välillä.

Mitä lyhyempi on kahden käsitteen välinen etäisyys, sitä läheisempiä ne ovat. Yksiselitteistämishypoteesi on, että kun on annettu joukko termejä, jotka esiintyvät lähellä toisiaan tekstissä ja joilla voi olla monia merkityksiä, oikeat merkitykset löydetään valitsemalla mahdollisista merkityksistä ne, jotka minimoivat etäisyydet.

Resnikin [Res95] esittämässä substantiiviryhmien WSD-menetelmässä yksiselitteistäminen suoritetaan joko suhteessa WordNetissa oleviin sanamerkityksiin tai korkeamman tason WordNet-kategorioihin. Korkeamman tason kategorioiden käyttö voi olla tarpeen, koska Resnikin mukaan WordNetin sanamerkitykset ovat niin hienojakoisia, että yksiselitteistäminen voi niiden tasolla olla vaikeaa. Resnikin algoritmi yksiselitteistää klusteroimalla tai manuaalisesti muodostettuja substantiiviryhmiä. Menetel-

mä käyttää pelkästään IS-A-suhteita. Yksiselitteistämisalgoritmin ydin on semanttisen samankaltaisuuden laskeminen käyttäen WordNetin taksonomiaa. Lähestymistavan takana oleva intuitio on seuraava: Mitä samankaltaisempia kaksi sanaa ovat, sitä informatiivisempi on jokin tietty yläkäsite (hyperonyymi), joka on molempien sanojen yläkäsite. Perinteinen tapa arvioida samankaltaisuutta semanttisessa verkossa on ollut mitata kahden solmun välisen polun pituus [RMBB89]. Käytännössä Resnikin lähestymistapa on samankaltainen. Jos IS-A suhteista muodostuva minimipolku kahden sanan välillä on pitkä, se tarkoittaa, että on tarpeen mennä korkealle taksonomiassa, abstraktimpiin käsitteisiin, jotta löydettäisiin yhteinen yläkäsite.

Substantiiviryhmien WSD-algoritmin perustana on oletus, että kun kaksi polyseemista sanaa ovat samankaltaisia, niiden informatiivisin yläkäsite tarjoaa informaatiota siitä, mikä kummankin sanan merkityksistä on relevantti. Sanat *doctor* ja *nurse* ovat polyseemisia. WordNetissa sanalla *doctor* on paitsi merkitys *health professional*, myös merkitys *someone who holds a Ph.D.* Sanalla *nurse* on paitsi merkitys *health professional*, myös merkitys *nanny*. Kun näitä kahta sanaa tarkastellaan yhdessä, kahden relevanteimman merkityksen yhteinen merkityskomponentti esiintyy informatiivisimman yläkäsitteen muodossa. Voi olla, että myös muilla mahdollisilla merkityksillä on jaettuja merkityskomponenetteja. Esimerkiksi *doctor, Ph.D, nurse* ja *nanny* ovat kaikki käsitteen {person, individual} alakäsitteitä. Mitä spesifisempi, tai informatiivisempi yhteinen esi-isä on, sitä varmemmin se osoittaa relevantit merkitykset, kun sanoja tarkastellaan yhdessä.

3.5 Yhteenveto

Tässä luvussa on tarkasteltu erilaisia WordNetin IS-A-hierarkiaa hyödyntäviä WSD-menetelmiä. Aqirren ja Rigaun [AR95, AR96], mukaan yksi WordNetiin liittyvä ongelma on sanaluokkien välisten suhteiden puuttuminen. Tälläiset suhteet saattaisivat

kuvata paremmin suhteita merkitysten välillä, ja tarjota näin paremman perustan yksiselitteistämiselle. Resnikin [Res95] mukaan WordNetin merkityserottelujen hienojakoisuus voi aiheuttaa ongelmia yksiselitteistämisessä. Resnikin esittämässä substantiiviryhmien WSD-menetelmässä ongelma on pyritty ratkaisemaan mahdollistamalla yksiselitteistäminen suhteessa korkeamman tason WordNet-kategorioihin.

Monet WordNetia ja tilastoja hyödyntävät WSD-menetelmät saavuttavat noin 60 prosentin tarkkuuden [MM98]. Mihalcean ja Moldovanin [RM00] esittämän WSD-menetelmän avulla voidaan yksiselitteistää vain ne sanat, joiden tapauksessa se voidaan tehdä melko varmasti oikein. Sanoista yksiselitteistetään noin 55 prosenttia, mutta 92 prosentin tarkkuudella. Tämä on hyödyllistä sovelluksissa, joissa yksiselitteistämisvirheistä on todennäköisesti enemmän haittaa kuin moniselitteisyydestä [San94].

4 Verbien aspektuaalinen yksiselitteistäminen

Tässä luvussa käsitellään Siegelin [Sie98] esittelemää menetelmää verbien aspektuaaliseen yksiselitteistämiseen WordNet-tietokannan avulla. Luvussa 4.1 tarkastellaan lyhyesti englannin kielen verbien aspektia. Luvussa 4.2 tarkastellaan aspektin suhteen moniselitteisiä verbejä. Luvussa 4.3 käsitellään Siegelin esittelemää menetelmää verbien aspektuaaliseen yksiselitteistämiseen niiden suoran objektin WordNet-kategorian perusteella.

4.1 Englannin kielen aspekti

Verbejä on jaoteltu esimerkiksi prosesseja, tiloja, mielenlaatua, tapahtumia ja saavutuksia kuvaaviin. Tapahtumien ajallista kestoa voidaan kuvata aikamuodon lisäksi muillakin kielellisillä keinoilla.

Aspektiluokka tarkoittaa sitä, että verbit ja verbiluokat eroavat toisistaan sen suhteen, millaisia maailman tiloja ne kuvaavat. Yleisesti oletetaan, että on olemassa kolme aspektiluokkaa: tila (state), toiminta (activity) ja tapahtuma (event) [Ven67, Pus95]. Tapahtuma-luokka voidaan edelleen jakaa suorituksiin (accomplishment) ja saavutuksiin (achievement).

Esimerkiksi *walked* virkkeessä 1 ilmaisee tapahtumaa, jonka kestoa ei ole määritelty. Tämä tarkoittaa, että virke ei sisällä tapahtuman ajallista kestoa koskevaa informaatiota, vaikka aikamuodon perusteella kyseessä on mennyt tapahtuma.

1) Mary *walked* yesterday

2) Mary *walked* to her house yesterday.

Virkkeen 1 kaltaisten virkkeiden sanotaan kuvaavan toimintaa [Ven67, Pus95]. Muita esimerkkejä toimintaa kuvaavista verbeistä ovat *run*, *sleep* ja *drink*. Virke 2 sisältää

saman informaation kuin virke 1 ja sen lisäinformaation, että Mary sai kävelemisensä päätökseen. Viittaamatta eksplisiittisesti tapahtuman ajalliseen kestoon, virke 2 antaa ymmärtää, että toiminnalla on looginen kulminaatiopiste, jossa se on ohi ja Mary on kotona. Tämäntyyppisen virkkeen sanotaan kuvaavan suoritusta.

Kuten verbi *walk* näyttää oletusarvoisesti kuvaavan toimintaa, jotkut verbit kuvaavat oletusarvoisesti suorituksia. Esimerkiksi *build* ja *destroy* tyypillisessä transitiivisessa käytössään kuvaavat suorituksia, koska suoritetuilla aktiviteeteilla on looginen kulminaatio:

3) Mary *built* a house.

4) Mary *destroyed* the table

Performanssiverbit, kuten *play*, voivat kuvata sekä toimintaa että suoritusta riippuen lauseen komplementtirakenteesta:

5) Mary *played* the piano.

6) Mary *played* the sonata in 15 minutes.

Kuten virke 6 osoittaa, yksi tapa testata, voiko verbi tai verbilauseke kuvata suoritusta, on sen modifiointi temporaalisella adverbiaalilla, kuten *in an hour*.

Saavutus on tapahtuma, joka aiheuttaa tilan muutoksen, kuten suorituskin, mutta muutoksen ajatellaan tapahtuvan hetkessä. Esimerkiksi virkkeissä 7, 8 ja 9 muutos ei ole asteittainen, vaan tapahtuu hetkessä. 'Piste-adverbiaalit', kuten *3 pm*, osoittavat, että virke kuvaa saavutusta.

7) John *died* at 3 pm.

8) John *found* his wallet at 3 pm.

9) Mary *arrived* at noon.

Piste-adverbien avulla voidaan kuitenkin modifioida myös suoritusta kuvaavia verbejä sisältäviä lauseita:

10) The pianist *performed* the sonata at noon.

11) James *taught* his 3 hour seminar at 2.30 pm. He *delivered* his lecture at 4:00 pm.

Näissä tapauksissa piste-adverbiaali osoittaa alkamisajan tapahtumalle, jolla on jokin tietty kesto.

Tilaa ilmaisevia (statiivisia) predikaatteja on kahdenlaisia: yksilötason (individual level) ja tilatason (stage level) predikaatteja. Predikaatteja *tall*, *intelligent* ja *owerweight* voidaan ajatella ominaisuuksina, jotka yksilö säilyttää mahdollisesti koko elämänsä ajan. Ne ovat yksilötason predikaatteja. Predikaatit *Hungry*, *sick* ja *clean* liittyvät tavallisesti ei-pysyviin tiloihin. Niitä kutsutaan tapahtumatason predikaateiksi. Tämä luokka esiintyy tyypillisesti kulminaatiota ilmaisevana predikaattina, kuten seuraavassa virkkeessä:

12) John drank himself *sick* with that cheap brandy.

Tiettyyn aspektiluokkaan kuuluminen määrittelee suurelta osin verbin semanttista käyttäytymistä. Verbin aspekti voi kuitenkin muuttua muiden tekijöiden tuloksena. Niitä ovat modifioivat adverbiaalit, argumenttina toimivan substantiivilausekkeen rakenne ja prepositiolausekkeet.

Aspektin mukainen luokittelu on pääkomponentti malleissa, jotka arvioivat temporaalisia rajoituksia lauseiden välillä [MS88, Sie98]. Esimerkiksi statiivisuus täytyy tunnistaa, jotta voitaisiin saada selville temporaaliset rajoitukset lauseiden välillä, joita yhdistää sana *when*. Seuraavassa esimerkissä temporaalinen suhde on tilan *have* ja tapahtuman *test* välinen:

13) She had a good strength when objectively tested.

Seuraavassa tapauksessa temporaalinen suhde taas on kahden tapahtuman välinen:

14) She had a seizure when objectively tested.

Tälläiset rajoitukset tarjoavat semanttisia rajoituksia luonnollisen kielen generoinnille ja ymmärtämiselle ja tarjoavat suuntaviivoja aspektuaaliselle korpusanalyysille.

4.2 Aspektin suhteen moniselitteiset verbit

Jotkut verbit näyttävät ilmaisevan vain yhtä aspektiluokkaa riippumatta kontekstista. Esimerkiksi *stare* (tuijottaa) ilmaisee ei-kulminoitunutta tapahtumaa. Monet verbit ovat kuitenkin aspektin suhteen moniselitteisiä. Esimerkiksi *show* kuvaa tilaa lauseessa *His lumbal puncture showed evidence of white cells*. Lauseessa *He showed me the photographs* se kuvaa tapahtumaa. Tämä moniselitteisyys vaikeuttaa automaattista verbin luokittelua, koska lauseen aspektiluokka riippuu pääverbin lisäksi useista muistakin lauseen konstituenteista [MS88].

Verbi *have* on erityisen ongelmallinen. Siegelin lääketieteellisessä aineistossa *have* esiintyi useiden lauseiden pääverbinä (8 prosentissa lauseista) [Sie98]. Noin 70 prosentissa lauseista se kuvasi tilaa ja 30 prosentissa tapahtumaa. Muilla moniselitteisillä verbeillä yksi merkitys oli dominoiva kyseisessä aihepiirissä.

Syntaktiset kategoriat, leksikaaliset pääsanat ja verbin argumenttien monikollisuus vaikuttavat aspektiluokkaan. Taulukko 5 havainnollistaa tätä. Taulukossa on esimerkkejä piirteistä, jotka vaikuttavat aspektiluokkaan. Kunkin piirteen vaikutusta havainnollistetaan esittämällä kaksi samantyyppistä lausetta, jotka kuuluvat eri aspektiluokkiin.

Aspektin suhteen moniselitteisten verbien sijoittaminen semanttisiin kategorioihin

Piirre	Esimerkki	luokka	vastaesimerkki	luokka
Predicate adj.	John *drove* the car	A	John *drove* the car ragged	E
Particle	John *drove* the car	A	John *drove* the car up	E
Dir object cat	John *saw* Sue	A	John *saw* that Sue was happy	E
Dir obj. head	Judith *played* the piano	A	Judith *played* the sonata	E
Dir obj. det	John *ate* fries	A	John *ate* the fries	E
Ind obj. det	Kathy *showed* people her car	A	Kathy *showed* the people her car	E
Ind obj. head	Kathy *showed* people her car	A	Kathy *showed* Sal her car	E
Prep obj head	Judith *looked* around the store	A	Judith *looked* around the corner	E
Prep obj det	Kathy *shot* at deer	A	Kathy *shot* at the deer	E
Tense	Sal *said* that it helps	E	Sal *says* that it helps	S

Taulukko 5: *Esimerkkipiirteet ja niiden vaikutus aspektiluokkaan. A tarkoittaa toimintaa, E tapahtumaa ja S tilaa.*

auttaa päättelemään, kuinka nämä verbit yhdessä argumenttiensa kanssa määrittelevät aspektiluokan. Tämä johtuu siitä, että monet verbit, joilla on samankaltainen merkitys toimivat argumenttiensa kanssa samalla tavoin. Yleensä verbin alikategorisaatiokehyksen (subcategorization frame) ja semanttisen luokan välillä on yhteys. Tämä pitää paikkansa erityisesti aspektin suhteen [Lev93]. Esimerkiksi *look* ja *weigh* voivat molemmat kuvata tapahtumia:

15) I looked at the baby.

16) I weighed the baby

Molemmat voivat kuvata myös tiloja:

17) The baby looked heavy.

18) The baby weighed a lot

Tämä havainnollistaa sitä, että näillä kahdella verbillä on samanlainen alikategorisaatiokehys, joka määrittelee niiden aspektiluokan. Myös niiden merkitysten välillä on korrelaatio, sillä kumpikin kuvaa jonkinlaista havaitsemista tai mittaamista.

Taulukko 6 kuvaa statiivin suhteen moniselitteisten verbien hierarkian huipun. Taulukossa esitetään seitsemän semanttista ryhmää esimerkkiverbien kanssa ja kaksi virkettä havainnollistaa esimerkkiverbin erilaisia käyttötapoja. Jokainen ensimmäisen ryhmän verbi voi kuvata joko tapahtumaa tai tilaa. Intuitiivisesti tämä johtuu siitä, että kukin verbi voi ilmaista kommunikatiivista tekoa.

Perception ja *psych-movement* ryhmät ovat ryhmän *cognition* alaryhmiä. Ryhmä *metaphorical* sisältää tapahtumaverbejä, joita käytetään idiomaattisesti ja jotka ovat statiivisia. Idiomaattinen käyttö vastaa tapahtuman metaforista selitystä. Esimerkiksi:

19) I ran down the street.

Ryhmä	Esimerkkiverbejä	Tapahtumavirke	Tilavirke
communication	*admit, confirm, indicate, say*	I *said* "Hello."	I *say* it is correct
cognition	*judge, remember, think, wish*	I *thought* about them	I *think* they are nice
perception	*feel, see, smell, weigh*	I *felt* the tablecloth	I *felt* terrible
psych-movement	*astonish, dismay, please, suprise*	You *suprised* me	That *suprises* me
location	*hold, lie, sit, stand*	I *lay* on the bed	The book *lies* on the bed
metaphorical	*work, run*	I *worked* hard	The machine *works*
carrier	*continue, remain*	I *continued* to talk about it	I *continued* to feel good

Taulukko 6: *Statiivin suhteen moniselitteisten verbien ryhmät*

20) It runs in the family.

Carrier-verbit heijastavat yksinkertaisesti lauseargumenttiensa aspektiluokkaa.

4.3 WordNetin käyttö aspektiluokan määrittelyssä

Siegel [Sie98] esittää säännön *have*-lauseiden luokitteluun sen mukaan, mihin kategoriaan niiden suorat objektit WordNetin perusteella kuuluvat. Säännön muodostamisessa huomioidaan seuraavat tekijät:

1) *Have*-sanan objektien jakaumat korpuksessa.

2) WordNet-kategorioiden ja aspektiluokan mukainen kielellinen intuitio.

3) Korrelaatiot WordNetin suoran objektin kategorian ja statiivisuuden välillä ohjatussa oppimisaineistossa.

Tämän informaation keräämiseksi suoritettiin WordNet-kysely jokaiselle jäsennetyssä korpuksessa esiintyvälle suoralle objektille. Kukin substantiivi sijoitettiin johonkin WordNetin semanttisen hierarkian huipulla olevista 25 kategoriasta. Monilla substantiiveilla on monia yksiköitä WordNetissa, koska niillä on monia merkityksiä. Aluksi otetaan käsiteltäväksi ensimmäinen lueteltu WordNet-kategoria, eli substantiivin yleisin merkitys. Pronominit, kuten *it* ja *him*, liitetään omaan pronominikategoriaansa.

Kuten taulukosta 7 nähdään, *have*-sanan yleisimmät objektit aineistossa liittyvät nimenomaan lääketieteelliseen aihepiiriin. Taulukossa on sanan korkean tason semanttinen (WordNet) kategoria ja *have*-lauseiden luokittelu, kun kukin substantiivi on suorana objektina. WordNet kykenee käsittelemään tätä teknistä aihepiiriä, koska 89.1 prosentilla *have*-lauseista on suora objekti, joka on yleisesti tunnettu lääketieteellinen termi tai ei tekninen termi.

Taulukoissa 8 ja 9 on esitetty *have*-lauseiden luokittelu niiden suoran objektin semanttisen kategorian perusteella. Erityisesti, jos lauseiden suorat objektin kuuluvat

suora objekti	lukumäärä	WordNet- Kategoria	Lauseen luokka
history	624	time	state
episode	280	event	event
pain	192	cognition	state
fever	123	cognition	state
temperature	113	attribute	state
allergy	109	state	state
movement	106	act	event
course	96	act	event
<none>	91	<none>	state
sympton	81	cognition	state
complaint	73	state	state
seizure	72	event	event
nausea	67	cognition	state

Taulukko 7: *Sanan have yleiset objektit, niiden WordNet-kategoria ja have-lauseiden aspektiluokka.*

kategorioihin: *event, act, phonemenon, communication, possession* ja *food*, lause luokitellaan tapahtumaa kuvaavaksi, muuten se luokitellaan tilaa kuvaavaksi.

Suoran objektin kategoria	Lukumäärä	Yleisiä substantiiveja
act	1157	*movement(106), course(96), difficulty(66), scon(61), admission(60)*
event	655	*episode(280), seizure(72), pulse(28), recerrence(25), onset(24)*
phonemenon	242	*pressure(52), x-ray(30), flatus(21), response(19), intake(15)*
communication	194	*sign(25), resolution(22), effusion(18), section(17), electrocardiogram(12)*
possession	59	*loss(27), amount(15), residual(5), insurance(4), cut(3)*
food	17	*loss(27), amount(15), residual(5), insurance(4), cut(3)*

Taulukko 8: *Tapahtumaluokkaan kuuluvat have-lauseet. Lukumäärät on laskettu kaikista korpuksen have-lauseista.*

Jos *have*-sanan suora objekti kuvaa tapahtumaa, lause kuvaa tapahtumaa. Tästä syystä on selvää, miksi WordNet-kategoriat *event, act, phonemenon* ja *communication* osoittavat lauseen kuvaavan tapahtumaa. Nominalisoidut tapahtumaverbit, kuten *resolution* on sijoitettu näihin neljään kategoriaan WordNetista saadun informaation perusteella. Kategoria *possession* valittiin, koska useimmat tapaukset, joissa *possession* esiintyi *have*-sanan suorana objektina, ovat sanan *loss* ilmentymiä. Esimerkiksi *The patient had blood loss* kuvaa tapahtumaa. Kategoria *food* valittiin kattamaan idiomeja, kuten: *The patient had lunch.*

Suoran objektin kategoria	lukumäärä	Yleisiä substantiiveja
cognition	1146	*loss(27), amount(15), residual(5)*
state	875	*bun(5), coffee(2), vitamin(1)*
N/A	860	*pain(192), fever(123), sympton(81)*
time	636	*allergy(109), complaint(73), infection(56)*
artifact	415	*echocardiogram(51), hematocrit(41), ultrasound(34)*
attribute	349	*history(642), rhythm(8), past(3)*
entity	209	*catherer(20), stool(19), tube(17)*
measure	205	*temperature(113), shortness(46), tenderness(26)*
substance	182	*chest(20), head(13), abdomen(13)*
relation	116	*count(41), increase(18), bout(15)*
person	115	*blood(29), thallium(15), sodium(11)*
group	84	*change(40), rate(32), function(12)*
location	49	*child(13), aide(13), son(8)*
feeling	48	*culture(41), series(7), meeting(6)*
pronoun	39	*area(8), post(7), left(6)*
animal	12	*dog(3), paceer(2), pet(1)*

Taulukko 9: *Tilaluokkaan kuuluvat have-lauseet. Lukumäärät on laskettu kaikista korpuksen have-lauseista.*

Myös ohjattu oppimisaineisto tukee tätä luokittelusääntöä. Taulukossa 7 nähdään kunkin WordNet-kategorian osalta tapahtumaa ja tilaa kuvaavien *have*-lauseiden jakauma, kun niiden suora objekti kuuluu kyseiseen kategoriaan. Aspektin mukainen luokittelu on aihepiiristä riippuvainen ongelma. Vaikka verbien täydellinen aspektileksikko voi riittää luokittelemaan monet lauseet niiden pääverbin perusteella, verbin primäärinen luokka riippuu usein aihepiiristä. Esimerkiksi monissa aihepiireissä verbi *show* kuvaa primäärisesti tapahtumaa. Lääketieteellisessä aineistossa se kuitenkin kuvasi pääasiassa tiloja. Tämän takia on Siegelin mukaan tarpeen tuottaa leksikko erikseen kullekin aihepiirille [Sie98].

Kun sääntöä testattiin substantiivijoukossa, saavutettiin 79.6 prosentin tarkkuus.

5 WordNetin hyödyntäminen leksikaalisten ketjujen muodostamisessa

Tässä luvussa käsitellään lähestymistapoja leksikaalisten ketjujen muodostamiseen WordNet-tietokannan avulla. Luvussa 5.1 kuvataan lyhyesti leksikaalista koheesiota ja leksikaalisia ketjuja. Luvussa 5.2 käsitellään muutamia lähestymistapoja leksikaaliseen ketjutukseen. Luvussa 5.3 kuvataan yksityiskohtaisemmin Barzilayn ja Elhadadin [BE97] esittämää menetelmää leksikaalisten ketjujen muodostamiseen. Luvussa 5.4 tarkastellaan Hirstin ja St-Ongen esittämiä huomioita WordNetin soveltuvuudesta leksikaalisten ketjujen muodostamiseen [HSO98].

5.1 Leksikaalinen koheesio ja leksikaaliset ketjut

Hallidayn ja Hasanin [HH76] mukaan *leksikaalinen koheesio*:

> "Refers to relations of meaning that exist within the text, and that define it as a text."

Leksikaalinen koheesio luodaan käyttämällä leksikaalisesti kiinteitä suhteita. Saman sanan toisto, sanan synonyymin käyttö esim. *dog* ja *hound*, sanan hypernyymin käyttö esim. *car* ja *vehicle* (kulkuneuvo), ja kollokaatioiden käyttö, esim. *garden* (puutarha) ja *digging* (kaivaminen) ovat mahdollisia tapoja leksikaalisen koheesion luomiseen.

Koheesio on keino 'sitoa yhteen' tekstin osat. Koheesio saavutetaan semanttisesti läheisten termien käytöllä, viittauksilla, ellipseillä ja konjunktioilla [HH76]. Leksikaalinen koheesio luodaan käyttämällä toisiinsa semanttisessa suhteessa olevia sanoja. Halliday ja Hasan luokittelevat leksikaalisen koheesion toistokategoriaan ja kollokaatiokategoriaan. Toisto voidaan saavuttaa samojen sanojen toistolla tai synonyymien ja hyponyymien käytöllä. Kollokaatiosuhteet määrittelevät suhteen sanojen välillä,

joilla on taipumus esiintyä yhdessä samassa leksikaalisessa kontekstissa, esimerkiksi: *She works as a teacher at school.*

Diskurssin rakenteen ja koheesion välillä on läheinen yhteys. Toisiinsa suhteessa olevilla sanoilla on taipumus esiintyä yhdessä tekstin diskurssiyksikössä. Koheesio on yksi tekstin rakenteen merkeistä, ja leksikaalisia ketjuja voidaan käyttää sen tunnistamiseen. Muita merkkejä tekstin rakenteen havaitsemiseen ovat välimerkit, kappaleen merkitsimet ja aikamuodon muutokset.

Leksikaalisten ketjujen muodostaminen on prosessi, jossa tekstistä eristetään toisiinsa liittyvät sanat [Gre99b]. Yleensä teksti sisältää useita leksikaalisia ketjuja, joista jokainen kuvaa jonkin osan tekstin kohesiivisesta rakenteesta. Nämä ketjut voidaan muodostaa käyttämällä mitä tahansa leksikaalista resurssia, joka järjestää sanat niiden merkitysten mukaan. WordNetia ovat käyttäneet leksikaalisena resurssina Barzilay ja Elhadad [BE97], Green [Gre99a], Hirst ja Morris [JM91], Hirst ja StOnge [HSO98] ja Stairmand [Sta97].

Leksikaalisia ketjuja on hyödynnetty tiedonhaussa [Sta97], oikolukemisessa [HSO98], hypertekstilinkkien muodostamisessa [Gre99a] ja automaattisessa yhteenvetojen muodostamisessa tekstistä [BE97].

5.2 Lähestymistavat leksikaaliseen ketjutukseen

Leksikaalisten ketjujen eristämiseen tekstistä on useita mahdollisia algoritmeja. Stairmand [Sta97] käyttää yksinkertaista algoritmia ketjujoukon muodostamiseen. Ensiksi kootaan kaikki sanat tekstistä. Sitten jokaiselle sanalle generoidaan termijoukko sanoista, jotka ovat lähellä sitä WordNetissa. Ketjut rakennetaan sitten etsimällä yhteydet laajennettujen termijoukkojen välillä.

Hirstin ja St-Ongen [HSO98] ja Greenin [Gre99a] kuvaamat leksikaaliset ketjuttajat pyrkivät huomioimaan erivahvuiset sananmerkityssuhteet. Esimerkiksi sanan toisto

katsotaan vahvemmaksi suhteeksi kuin synonyymin käyttö. Suhteet sanojen välillä voidaan myös laskea huomioimalla tietynlaiset polut synonyymijoukkojen välillä WordNet-verkossa. Kun sanat luetaan tekstistä, ne liitetään ketjuun, jonka kanssa niillä on vahvin suhde.

Tälläisissä ketjuttajissa on se puute, että tieto siitä, millainen suhde ketjun sanojen välillä on hukataan heti, kun päätös lisätä sana ketjuun on tehty. Al-Halimin ja Kazmamin [AHK98] esittämässä lähestymistavassa muodostetaan leksikaalinen puu. Leksikaalinen puu säilyttää suhteet, jotka esiintyivät sanojen välillä ketjutusprosessin aikana.

Kaikki edellä kuvatuista menetelmistä ovat ahneita, eli päätös siitä, mihin ketjuun sana sijoitetaan tehdään niin nopeasti kuin mahdollista. Barzilayn ja Elhadadin [BE97] esittämä lähestymistapa on erilainen. Heidän leksikaalinen ketjuttajansa säilyttää kaikki mahdolliset selitykset, kunnes kaikki ketjutettavat sanat on huomioitu. Kaikkien mahdollisten selitysten säilyttäminen vaatii huomattavan määrän muistia. Barzilay ja Elhadad ratkaisivat tämän ongelman segmentoimalla tekstin, suorittamalla ketjutusalgoritmin kussakin segmentissä ja liittämällä sitten yhteen eri segmentteihin ulottuvat ketjut. Tästä huolimatta on luultavaa, että tämä algoritmi on hitaampi kuin edellä kuvatut [Gre99b].

Yksi leksikaalisen ketjutuksen hyödyllisimmistä ominaisuuksista on se, että sanamerkitysten yksiselitteistäminen tapahtuu prosessin sivuvaikutuksena [Gre99b]. Kun sanat liitetään leksikaalisiin ketjuihin, synonyymijoukot, joita ei voida liittää ketjussa jo oleviin synonyymijoukkoihin poistetaan. Leksikaalisten ketjujen jäsenet määrittelevät semanttisen kontekstin, jossa sanaa käytetään. Ja sanamerkitykset, jotka ovat mahdottomia tässä kontekstissa, poistetaan. On myös odotettavissa, että synonyymiset sanat esiintyvät samassa ketjussa, koska sellaiset sanat kuuluvat samaan synonyymijoukkoon.

5.3 Leksikaalisten ketjujen käyttö yhteenvetojen tuottamisessa

Barzilay ja Elhadad [BE97] tutkivat tekniikkaa yhteenvedon tuottamiseksi alkuperäisestä tekstistä, ilman että tarvitaan tekstin täyttä semanttista analyysia. Tekniikka nojaa leksikaalisten ketjujen avulla derivoituun aihepiiriin. Barzilay ja Elhadad esittävät algoritmin leksikaalisten ketjujen derivointiin tekstistä. Algoritmi perustuu WordNet-tietokantaan, sanaluokkajäsentimeen, substantiiviryhmiä tunnistavaan jäsentimeen ja Hearstin [Hea94] esittämään segmentointialgoritmiin. Yhteenvedon muodostaminen tapahtuu kolmessa askeleessa:

1) Alkuperäinen teksti segmentoidaan.

2) Leksikaaliset ketjut muodostetaan.

3) 'Vahvat' ketjut tunnistetaan, ja tärkeät virkkeet eristetään tekstistä.

Yleisesti ottaen leksikaalisten ketjujen muodostaminen tapahtuu kolmessa vaiheessa:

1) Valitaan joukko kandidaattisanoja.

2) Jokaiselle kandidaattisanalle etsitään sopiva ketju. Ellei sellaista löydy, sille luodaan uusi ketju.

3) Jos sopiva ketju löytyy, sana lisätään siihen ja ketju päivitetään.

Barzilay ja Elhadad [BE97, HSO98] esittävät seuraavan esimerkin tästä proseduurista: Esiprosessointivaiheessa valitaan kaikki sanat, jotka esiintyvät substantiiveina WordNetissa. Sanojen välinen suhde määritellään sen perusteella, mikä on niiden välinen etäisyys tekstissä, ja sen perusteella, millainen polku yhdistää niitä WordNetissa. Suhteita on kolmenlaisia: hyvin vahva (extra-strong) kahden saman sanan esiintymän välillä, vahva (strong) kahden sanan välillä, joita yhdistää WordNet-linkki, ja keskivahva (medium-strong), kun sanojen esiintymissynonyymijoukkoja yhdistävän polun pituus on suurempi kuin yksi. Vain tiettyjen rajoitusten mukaiset polut hyväksytään

valideiksi linkeiksi. Maksimietäisyys toisiinsa suhteessa olevien sanojen välillä riippuu suhteen tyypistä. Hyvin vahvan suhteen tapauksessa etäisyyttä ei ole rajoitettu. Vahvan suhteen tapauksessa se on rajoitettu seitsemän virkkeen kokoiseen ikkunaan. keskivahvan suhteen tapauksessa se ulottuu kolme virkettä taaksepäin.

Etsittäessä ketjua, johon lisätä kandidaattisana, hyvin vahvoilla suhteilla on etusija vahvoihin suhteisiin verrattuna, ja molemmilla on etusija keskivahvoihin suhteisiin verrattuna. Jos sopiva ketju löydetään, kandidaattisana lisätään ketjuun sopivimmassa merkityksessä ja muut ketjussa olevat sanat päivitetään, niin että jokainen ketjussa oleva, uuteen sanaan liittyvä sana on suhteessa sen valittuun merkitykseen. Jos sopivaa ketjua ei löydetä, luodaan uusi ketju, johon uusi sana lisätään sen kaikissa WordNetin mukaisissa merkityksissä.

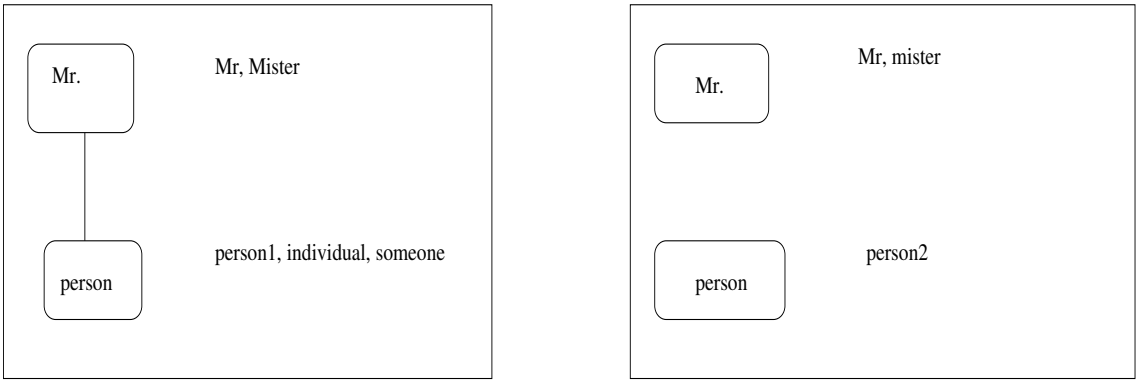
Havainnollistetaan yksiselitteistämistä seuraavalla esimerkillä [BE97] :

> "Mr Kennedy is the person, that invented an anaesthetic machine which uses micro-computers to control the rate at which an anaesthetic is pumped into the blood. Such machines are nothing new but his device uses two micro-computers to achieve much closer monitoring of the pump feeding the anaesthetic into the patient."

Ensiksi luodaan solmu sanalle *Mr* {mister, Mr}. Seuraava kanditaattisana on *person*. Sillä on kaksi merkitystä WordNetissa: *human being* ja *grammatical category of pronouns and verb forms*. Sanan *person* merkityksen valinta jakaa ketjun kahteen erilaiseen selitykseen, kuten kuvassa 2 on esitetty.

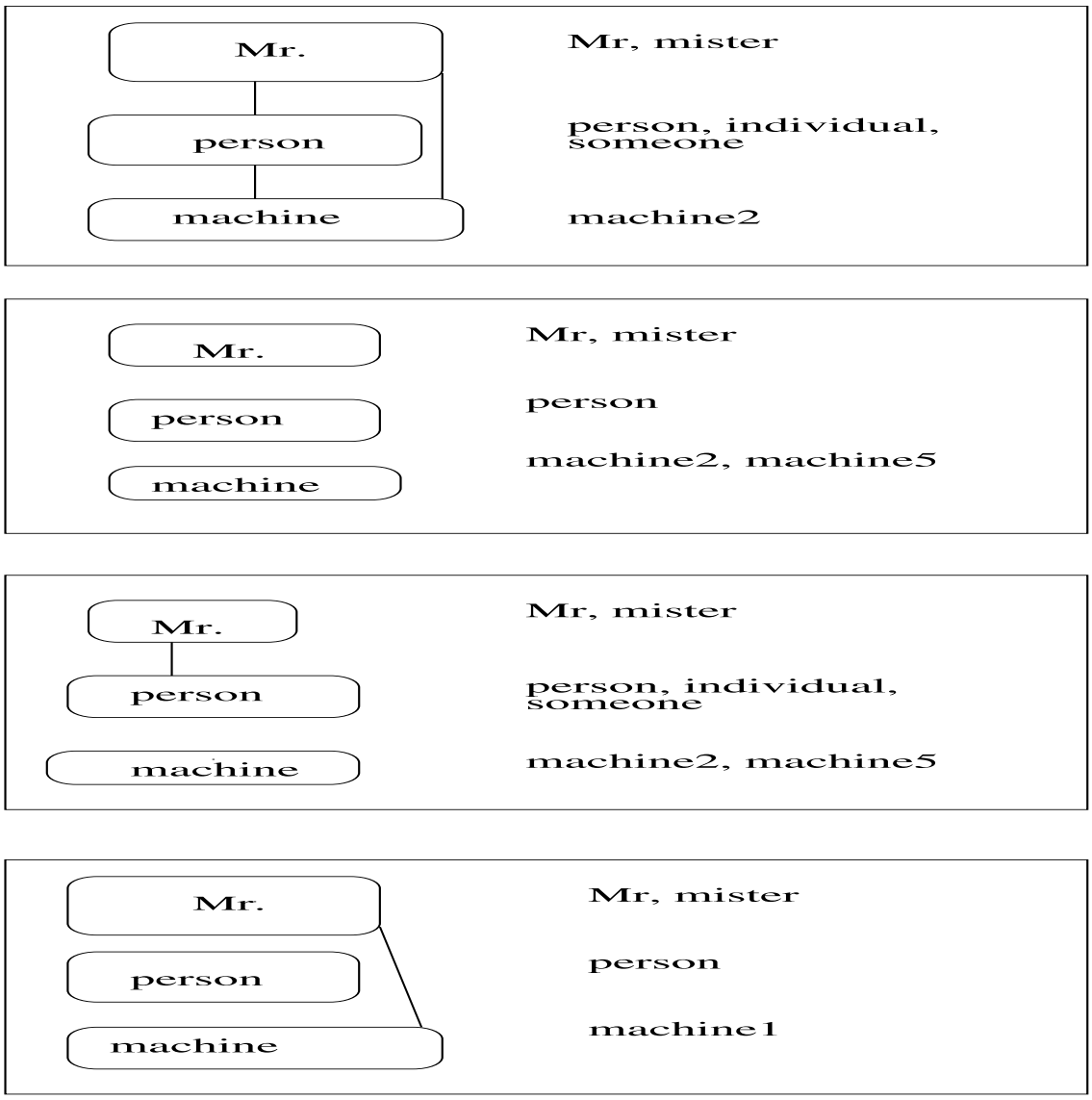
Komponentti määritellään listaksi toisensa poissulkevia selityksiä. Komponenttisanat vaikuttavat toisiinsa niiden merkityksiä valittaessa.

Seuraava kandidaattisana *anaesthetic* ei ole suhteessa ensimmäisen komponentin sanoihin, joten sille luodaan uusi komponentti.



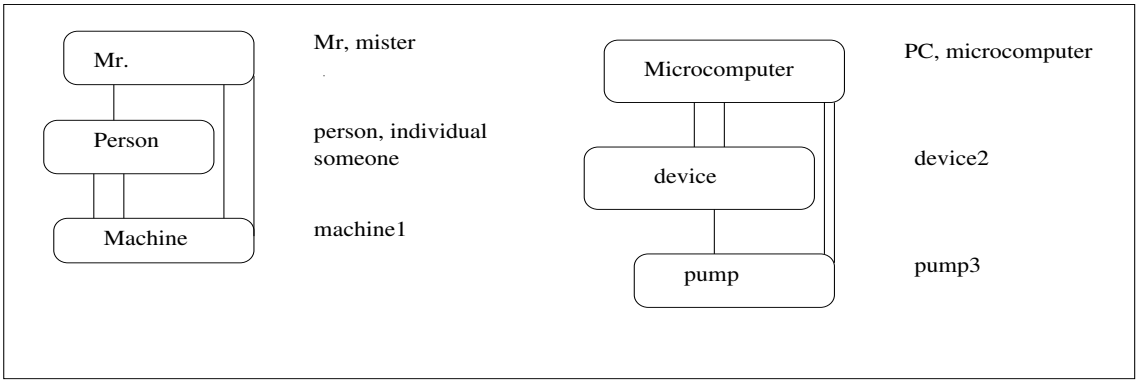
Kuva 2: Askel 1, selitykset 1 ja 2

Sanalla *machine* on viisi merkitystä. Ensimmäisessä merkityksessään *an efficient person*, se on suhteessa merkityksiin *person* ja *Mr.*. Se vaikuttaa niiden merkitysten valintaan. *Machine* lisätään ensimmäiseen komponenttiin. Lisäyksen jälkeen ensimmäisen komponentin sisältö on kuvassa 3 esitetyn kaltainen. Lisättäessä sanat *microcomputer*, *device* ja *pump* vaihtoehtojen määrä kasvaa. Todennäköisimmät selitykset on esitetty kuvissa 4 ja 5.

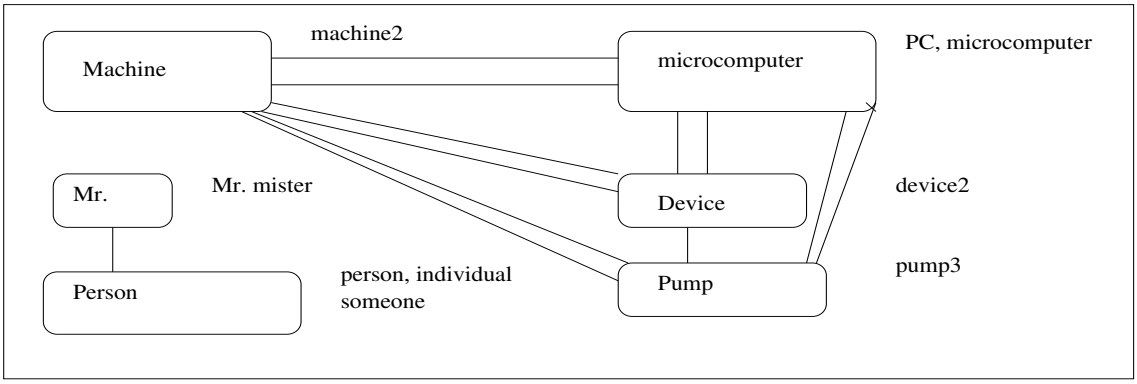


Kuva 3: Askel 2, selitykset 1, 2, 3 ja 4

Oletetaan, että teksti on kohesiivinen, ja määritellään paras selitys selitykseksi, jolla on eniten yhteyksiä (kaaria verkossa). Tässä tapauksessa valitaan toinen selitys askeleen kolme lopussa, joka ennustaa oikean merkityksen sanalle *machine*. Selityksen todennäköisyysarvo määritellään sen ketjun todennäköisyysarvojen summaksi. Ketjun todennäköisyysarvo määritellään sen jäsenten välisten suhteiden määrän ja painon mukaan. Toiston ja synonyymin painoksi annetaan 10, antonyymin painoksi 7 ja



Kuva 4: Askel 3, selitys 1



Kuva 5: Askel 3, selitys 2

hyperonyymin ja holonyymin painoksi 4. Algoritmi säilyttää kaikki mahdolliset selitykset. Kun mahdollisten selitysten määrä on tiettyä kynnysarvoa suurempi, heikot selitykset karsitaan. Lopuksi jokaisesta komponentista valitaan vahvin selitys.

Useimmat leksikaalisten ketjujen muodostamiseen käytettävät menetelmät valitsevat kanditaattisanoiksi vain substantiiveja [HSO98]. Barzilay ja Elhadad valitsevat myös *substantiiviyhdistelmiä* (noun compound). Noin 50 000 WordNetin yksikköä on substantiiviyhdistelmiä, kuten *sea level* tai kollokaatioita, kuten *digital computer*. Englannin kieli on kuitenkin produktiivinen substantiiviyhdistelmien suhteen. Jokaisessa aihepiirissä esiintyy uusia substantiiviyhdistelmiä ja kollokaatioita, joita ei ole WordNetissa.

Barzilay ja Elhadad käyttävät pinnallista jäsennystä substantiiviyhdistelmien tunnistamiseen. Käytössä on yksinkertainen substantiiviyhdistelmien kuvaus. Substantiiviyhdistelmien tunnistaminen on kahdella tavalla hyödyllistä.

1) Sen avulla tunnistetaan aihepiirin tärkeitä käsitteitä esimerkiksi *quantum computing*, joka ei esiinny WordNetissa.

2) Sen avulla voidaan eliminoida substantiiviyhdistelmissä modifioijina esiintyvät sanat, eikä niitä pidetä mahdollisina ketjun jäseninä. Esimerkiksi, kun *quantum computing* käsitetään yhdeksi yksiköksi, sanaa *quantum* ei valita kandidaattisanaksi. Tämä on hyödyllistä, jos dokumentin aiheena ei ole *quantum* (kvantti) vaan *computing*.

5.4 WordNet tiedon lähteenä leksikaaliselle ketjuttajalle

Hirstin ja St-Ongen mukaan se, että WordNetissa ei ole linkkejä substantiivitietokannasta muihin tietokantoihin estää käyttämästä muita sanoja leksikaalisten ketjujen muodostamiseen [HSO98].

Leksikaalisten ketjujen muodostamisessa voi tapahtua kahden tyyppisiä virheitä; Sanoja ei lisätä ketjuun, johon ne kuuluvat, tai niitä lisätään ketjuun, johon ne eivät kuulu. Hirstin ja St-Ongen [HSO98] mukaan näitä virheitä voi aiheutua mm. seuraavista syistä:

1) Rajoitukset WordNetin suhteiden joukossa, tai puuttuvat linkit.

2) Implisiittinen epäjohdonmukaisuus semanttisen etäisyyden suhteen WordNet-suhteissa.

3) Virheellinen tai epätäydellinen yksiselitteistäminen.

4) Sanojen metaforinen käyttö.

Seuraava virke antaa esimerkin ensimmäistä tyyppiä olevasta ongelmasta:

> "School administration say these sama taxpayers expect the *schools* to provide *child care* and *school* lunches, to integrate immigrants into the community, to offer special classes for adult students,... "

Tässä tapauksessa haluttaisiin saada ilmaus *child care* liitetyksi sanaan *school*. WordNetissa ei kuitenkaan ole riittävää määrää suhteita näiden sanojen yhdistämiseen. Näiden kahden sanan välinen suhde on tilannekohtainen.

Toista ongelmatyyppiä havainnollistaa seuraava virke:

> "The cost means no holiday trips and more *stew* than *steak*, but she is satisfied that her children, now in grades 3 and 4, are being properly taught. "

Sanat *stew* ja *steak* ovat semanttisessa suhteessa toisiinsa. WordNetissa niitä ei ole kuitenkaan liitetty toisiinsa. Niiden suhde on määritelty WordNetissa seuraavasti:

stew IS A *dish* IS A aliment INCLUDES meat INCLUDES cut / cut of mead INCLUDES piece / slice INCLUDES steak

Etäisyys sanat *stew* ja *steak* sisältävien synonyymijoukkojen välillä on kuusi synonyymijoukkoa, mikä on enemmän kuin leksikaalisessa ketjuttajassa asetettu raja.

On myös tilanteita, joissa WordNetissa lähellä toisiaan olevat sanat ovat semanttisesti melko etäällä toisistaan. Tämä voi aiheuttaa sanan liittämisen ketjuun, johon se ei kuulu. Esimerkiksi eräässä ketjussa sana *public* oli liitetty sanaan *professional* seuraavan suhteen perusteella:

public IS A *people* HAS MEMBER *person* INCLUDES *adult* INCLUDES *professional*

Virheellisen yksiselitteistämisen ongelma aiheutuu usein ali- tai yliketjutuksesta:

> "We suppose a very long *train* traveling along the rails with the constant *velocity v* and in the direction indicated..."

Tässä tapauksessa on luotu ketju sanalle *train*, jolla on kuusi merkitystä WordNetissa. Sanaa *rails* ei kuitenkaan liitetä siihen, koska sanojen välinen etäisyys WordNetissa on liian suuri. Sana *velocity* liitetään sitten sanaan *train*. Tämän takia sanalle *train* valitaan väärä merkitys:

sequence, succession, sequel, train-events that are ordered in time

6 WordNet ja tiedonhaku

Tässä luvussa käsitellään WordNet-tietokannan hyödyntämistä tiedonhaussa. Luvussa 6.1 tarkastellaan yksiselitteistämisen vaikutusta tiedonhakuun. Luvussa 6.2 käsitellään Mihalcean ja Moldovanin [MM00] esittämää tiedonhakujärjestelmää, joka hyödyntää sekä semanttista indeksointia, että leksikaalisiin sanoihin perustuvaa indeksointia. Luvussa 6.3 käsitellään leksikaalisten ketjujen hyödyntämistä indeksoinnissa ja dokumenttien haussa [Sta97].

Tiedonhaussa on tarkoitus löytää kaikki relevantit dokumentit dokumenttikokoelmasta kyselyn perusteella [GVCC98]. Monet tiedonhakujärjestelmät perustavat arvionsa dokumenttien ja kyselyiden samankaltaisuudesta siihen, kuinka monta yhteistä sanaa niissä on. Mitä enemmän dokumentissa ja kyselyssä on yhteisiä sanoja, sitä relevantimmaksi dokumentti katsotaan. Järjestelmän suorituskykyä voidaan parantaa painottamalla kyselyn ja dokumentin sanoja käyttäen dokumenttikokoelmasta ja yksittäisestä dokumentista saatavaa frekvenssi-informaatiota. Dokumentit palautetaan tyypillisesti järjestettynä listana, jossa järjestys perustuu arvioon dokumentin relevanssista [KC92].

Pääongelma leksikaalisiin sanoihin perustuvassa tiedonhaussa on se, että se palauttaa tavallisesti liian monia dokumentteja ja/tai vääriä dokumentteja. Avainsanoilla voi olla useita semanttisia merkityksiä. Sanojen eri merkitykset voivat myös kuulua eri sanaluokkiin [MM00]. Toisaalta joitakin relevantteja dokumentteja voi jäädä löytymättä, koska dokumentissa saatetaan käyttää eri termejä kuin kyselyssä.

WordNetia on hyödynnetty dokumenttien haussa kolmella tavalla. Dokumentteja ja kyselyitä on yksiselitteistetty, kyselyitä on rikastettu semanttisesti läheisillä termeillä ja kysymyksiä ja dokumentteja on vertailtu käsitteellisen etäisyyden perusteella.

Kyselyn rikastamisen WordNetin avulla on osoitettu olevan potentiaalisesti hyödyllistä [SQ96]. Voorhees [Voo94] laajensi manuaalisesti 50 TREC-1 kokoelmassa esiintyvää

kyselyä, käyttäen synonymiaa ja muita WordNetin semanttisia suhteita. Voorheesin mukaan laajentaminen oli hyödyllistä lyhyiden, epätäydellisten kyselyiden tapauksissa, mutta melko hyödytöntä täydellisempien kyselyiden kohdalla, jolloin muut laajennustekniikat toimivat paremmin. Lyhyiden kyselyjen tapauksessa jäi ongelmaksi, kuinka laajennukset valitaan automaattisesti. Sen tekeminen huonosti voi huonontaa hakutulosta parantamisen sijaan. Richarson ja Smeaton [RS97] käyttivät WordNetiin perustuvien tekniikoiden yhdistelmää, mukaanlukien automaattinen WSD ja semanttisen etäisyyden mittaaminen kyselyn ja dokumentin termien välillä. Tämä johti tehokkuuden huononemiseen leksikaalisiin sanoihin perustuvaan hakuun verrattuna. Smeaton ja Quigley [SQ96] hyödynsivät semanttisen etäisyyden mittaamista dokumenttien ja kyselyiden välillä suorittaessaan hakuja pienestä kokoelmasta kuvatekstejä. Tämä kokoelma sisälsi hyvin lyhyitä dokumentteja. Tässä tapauksessa tehokkuus parani siksi, että todennäköisyys löytää alkuperäiset kyselytermit sisältävä dokumentti on paljon pienempi kuin pidempien dokumenttien tapauksessa, jotka usein sisältävät samoja käsitteitä ilmaistuna monilla tavoilla.

6.1 Yksiselitteistäminen ja tiedonhaku

Käsitykset WSD:n hyödyllisyydestä tiedonhaussa ovat osittain ristiriitaisia. Sandersonin [San94] mukaan ratkaisematon moniselitteisyys ei juuri vaikuta tiedonhakujärjestelmien suorituskykyyn, mutta yksiselitteistämisessä tapahtuvat virheet sen sijaan vaikuttavat.

Myöskään Krovetzin ja Croftin mukaan yksiselitteistäminen ei ole yleisesti ottaen hyödyllistä eroteltaessa relevantteja ja ei-relevantteja dokumentteja [KC92]. Heidän suorittamissaan testeissä vain 7-13 prosentissa hakujärjestelmän palauttamista dokumenteista tapahtui merkitysten sekaantumisia, eli hakutermi esiintyi dokumentissa eri merkityksessä kuin kyselyssä. Heidän mukaansa WSD voi kuitenkin olla joissakin

tapauksissa hyödyllistä, mikäli löydetään tapa tunnistaa ne sanat, jotka kannattaa yksiselitteistää. Sanat, jotka kannattaa yksiselitteistää, jakaantuvat kahteen kategoriaan:

1) Sanalla ei ole yhtä yleistä merkitystä, joka kattaisi yli 80 prosenttia sen esiintymistä.

2) Sanalla on yksi yleinen merkitys, mutta kyselyssä esiintyvä merkitys on yksi harvinaisemmista.

Krovetzin ja Croftin mukaan tiettyihin aihepiireihin liittyvissä dokumenttikokoelmissa sanojen moniselitteisyys on todennäköisesti pienempi ongelma kuin heterogeenisissa dokumenttikokoelmissa, koska moniselitteisillä sanoilla on todennäköisesti yhdenmukainen merkitys tietyssä aihepiirissä. Heterogeenisissa dokumenttikokoelmissa yksiselitteistäminen on kenties vaivan arvoista.

Schutzen ja Pedersenin [SP95] mukaan haku yksiselitteistetyillä avainsanoilla voi kasvattaa tiedonhakujärjestelmän tarkkuutta 7 prosenttia. Merkitykseen perustuvan ja sanamuotoon perustuvan haun yhdistelmä voi kasvattaa tarkkuutta 14 prosenttia.

Gonzalon et. al [GVCC98] mukaan yksiselitteistäminen parantaa tiedonhakujärjestelmän suorituskykyä jos korkeintaan 30 prosenttia kohdesanoista on yksiselitteistetty virheellisesti. Kun yksiselitteistämisvirheiden osuus on 30-60 prosenttia, yksiselitteistäminen ei edelleenkään huononna hakutulosta.

6.2 Yhdistetty semanttinen ja sanamuotoon perustuva indeksointi

Mihalcea ja Moldovan [MM00] esittelevät tiedonhakujärjestelmän, joka liittää sanasemantiikkaa klassiseen sanamuotoon perustuvaan indeksointiin. Järjestelmän kaksi pääkomponenttia, indeksointi- ja hakukomponentti käyttävät yhdistettyä sanaan perustuvaa ja merkitykseen perustuvaa lähestymistapaa. Järjestelmä perustuu meto-

dologiaan, jonka avulla vapaasta tekstistä rakennetaan semanttisia representaatioita sana-, ja kollokaatiotasolla. Tämä semanttiseksi indeksoinniksi kutsuttu tekniikka on osoittautunut heidän mukaansa tehokkaammaksi kuin pelkästään sanamuotoon perustuva indeksointi.

Pääongelma perinteisessä sanaan perustuvassa tiedon haussa on se, että se palauttaa tavallisesti liian monia tuloksia ja/tai vääriä tuloksia.

Ratkaisu on sisällyttää enemmän informaatiota indeksoitaviin dokumentteihin, esimerkiksi mahdollistaa se, että järjestelmä voi hakea dokumentteja joko perustuen sanoihin (leksikaalisiin merkkijonoihin), tai perustuen sanojen semanttisiin merkityksiin.

Näiden ideoiden perusteella Mihalcea ja Moldovan ovat suunnitelleet tiedonhakujärjestelmän, joka sovittaa yhteen sanamuotoon perustuvan ja merkitykseen perustuvan indeksoinnin ja haun.

Järjestelmän syöte koostuu kyselystä ja joukosta dokumentteja, joista tieto haetaan. Leksikaalinen ja semanttinen informaatio lisätään sekä kyselyyn että dokumentteihin esiprosessointivaiheen aikana, jossa kysely ja tekstit yksiselitteistetään. WSD-proseduuri nojaa konteksti-informaatioon ja tunnistaa sanamerkitykset WordNet-merkitysten perusteella. Käytetty WSD-algoritmi yksiselitteistää noin 55 prosenttia substantiiveista ja verbeistä, mutta se on hyvin tarkka (noin 92 prosentin tarkkuus) [RM00]. Jokaiseen sanaan liitetään myös sanaluokkatunniste. Kun nämä leksikaaliset ja semanttiset tyypit on liitetty sanoihin, dokumentit ovat valmiita indeksoitaviksi. Indeksi luodaan käyttäen sanoja leksikaalisina merkkijonoina ja semanttisia tunnisteita.

Kun indeksi on luotu, syötekyselyyn vastataan käyttäen järjestelmän dokumenttienhakukomponenttia. Ensiksi kysely yksiselitteistetään. Sitten se muunnetaan erityiseen esitysmuotoon, joka liittää yhteen indeksistä löytyvää semanttista informaatiota.

Semanttisen indeksoinnin käytöllä pyritään ratkaisemaan kaksi tiedonhakujärjestelmiin liittyvää pääongelmaa:

1) Relevanttia informaatiota ei hukata olemalla spesifioimatta avainsanoja. Kun sanoihin liitetään uusia merkitystunnisteita, voidaan hakea myös sanoilla, jotka ovat semanttisesti lähellä avainsanoja.

2) Käyttämällä hakujärjestelmän merkitykseen perustuvaa komponenttia hakutuloksia voidaan vähentää spesifioimalla tarkkaan syöteavainsanan leksikaalinen funktio, ja/tai merkitys.

Järjestelmää testattiin Cranfieldin standartitekstikokoelman avulla. (Cranfieldin standartitekstikokoelma koostuu 1400:sta SGML muotoisesta dokumentista. Dokumentit käsittelelevät airodynamiikkaa.) Jokaisesta testatusta 50 kysymyksestä muodostettiin kolme erityyppistä kyselyä:

1) Kysely, joka käyttää ainoastaan kysymyksestä valittuja avainsanoja.

2) Kysely, joka käyttää kysymyksessä esiintyviä avainsanoja, ja näiden avainsanojen synonyymijoukkoja.

3) Kysely, joka käyttää kysymyksessä esiintyviä avainsanoja, avainsanojen synonyymijoukkoja ja avainsanojen hypernyymien synonyymijoukkoja.

Kaikkia näitä kyselyitä on testattu semanttisen indeksin avulla. Tulokset osoittavat, että suorituskyky paranee, kun hakujärjestelmä käyttää sekä sanamuotoon että merkitykseen perustuvaa indeksointia, verrattuna klassiseen sanamuotoon perustuvaan indeksointiin.

6.2.1 Järjestelmän arkkitehtuuri

Mihalcean ja Moldovanin esittämässä järjestelmässä pyritään yhdistämään sanamuotoon perustuvan ja synonyymijoukkoon perustuvan indeksoinnin edut. Sekä sanat

että synonyymijoukot on indeksoitu syötetekstissä ja haku suoritetaan käyttäen jompaakumpaa, tai molempia näistä informaation lähteistä.

Järjestelmässä on kolme päämoduulia:

1) WSD-moduuli, joka suorittaa osittaisen, mutta tarkan yksiselitteistämisen. Semanttisen informaation lisäksi se lisää sanaluokkatagit jokaiseen sanaan ja etsii sanan vartalon WordNetin morphword funktion avulla. Jokainen dokumentti prosessoidaan tämän moduulin avulla. Tulos on uusi dokumentti, jossa jokainen sana on korvattu rakenteella:

Pos|Stem|POS|Offset

missä Pos on sanan sijainti tekstissä, Stem on sanan vartalo, POS on sanaluokka ja Offset on sen WordNetin synonyymijoukon Offset, jossa sana esiintyy. Tapauksissa, joissa WSD-moduuli ei ole liittänyt sanaan mitään merkitystä, tai sanaa ei löydy WordNetista, viimeinen kenttä jätetään tyhjäksi.

Käytetty WSD-algoritmi on iteratiivinen. Se määrittää, mitkä annetun tekstin substantiiveista ja verbeistä voidaan yksiselitteistää tarkasti. Semanttinen luokittelu suoritetaan käyttäen WordNetissa määriteltyjä merkityksiä [RM00].

2) Indeksointimoduuli, joka indeksoi dokumentit, kun WSD-moduuli on prosessoinut ne. WSD-funktion palauttamasta sanan uudesta rakenteesta Stem ja Offset/POS liitetään erikseen indeksiin. Tämä mahdollistaa haun sanamuodoilla tai sanojen synonyymijoukoilla.

3) Hakumoduuli, joka hakee dokumentteja syötekyselyn perusteella. Voidaan hakea joko avainsanoja sisältäviä dokumentteja, dokumentteja, jotka sisältävät avainsanoja, joihin on liitetty merkitys, tai dokumentteja, jotka sisältävät avainsanojen synonyymeja.

6.2.2 Tulokset

Järjestelmän suorituskyvyn arviointiin on käytetty kolmea mittaria:

1) Tarkkuus. Relevanttien dokumenttien osuus kaikista palautetuista dokumenteista.

2) Saanti. Palautettujen relevanttien dokumenttien osuus kaikista kokoelmassa olevista relevanteista dokumenteista.

3) F-mitta, joka yhdistää tarkkuuden ja saannin:

$$\mathrm{F-measure} = \frac{(B^2 + 1.0) * P * R}{(B2 * P) + R}$$

Kaavassa P on tarkkuus, R on saanti ja B on suhteellinen tärkeys, joka tarkkuudelle annetaan suhteessa saantiin.

Testattaessa järjestelmää kaikilla 50 kysymyksellä saatiin seuraavat tulokset:

Tarkkuus oli 0.22 ja saanti 0.25 käytettäessä sanamuotoon perustuvaa indeksointia. Tarkkuus oli 0.23 ja saanti 0.29 käytettäessä yhdistettyä sanamuotoon perustuvaa ja synonyymijoukkoon perustuvaa indeksointia. Tarkkuus oli 0.21 ja saanti 0.32 käytettäessä hypernyymisynonyymijoukkoja.

Yhdistetyllä sanamuotoon ja synonyymijoukkoon perustuvalla indeksoinnilla saavutettu etu on 16 prosentin kasvu saannissa ja 4 prosentin kasvu tarkkuudessa, verrattuna sanamuotoon perustuvaan perusindeksointiin.

Kuinka tehokas tälläinen järjestelmä on? Tiedonhakuprosessiin on liitetty WSD-proseduuri ja tiedetään, että WSD-algoritmit ovat yleensä laskennallisesti vaativia. Toisaalta dokumenttikokoelman yksiselitteistäminen on prosessi, joka voidaan suorittaa suurelta osin rinnakkaisesti, eikä se näin ollen muodostu ongelmaksi.

6.3 Leksikaalisten ketjujen hyödyntäminen indeksoinnissa

Stairmand [Sta97] kuvaa leksikaalisiin ketjuihin perustuvan indeksointikomponentin tiedonhakujärjestelmille. Järjestelmän kaksi perusrakennetta ovat *leksikaalinen klusteri* ja *leksikaalinen ketju*.

Leksikaalinen klusteri koostuu sanaston yksiköistä, jotka muodostavat erillisen tekstuaalisen kontekstin. Tekstuaalinen konteksti voi olla aktiivinen dokumentin eri osissa. Leksikaalisen ketjun avulla voidaan kuvata, missä osissa dokumenttia tekstuaalinen konteksti on aktiivinen. Tämä auttaa puolestaan eliminoimaan väärät sanat, koska klusterin validien sanojen täytyy kuulua myös leksikaaliseen ketjuun. Leksikaalinen klusteri koostuu yhdestä tai useammasta leksikaalisesta ketjusta. Taulukossa 10 on esitetty kolme erilaista leksikaalista kontekstia ja vastaavat leksikaaliset ketjut, jotka osoittavat, missä osissa dokumenttia konteksti on aktiivinen (esim. $Appeal_{1.1}$ osoittaa, että termi *appeal* tässä asemassa kuuluu klusteriin 1, ketjuun 1). Leksikaalisten ketjujen muodostamista on kuvattu tarkemmin luvussa 5.

Stairmandin hypoteesin mukaan sanojen esiintymiskontekstin hyödyntäminen indeksoinnissa ja dokumenttien haussa parantaa suorituskykyä verrattuna pelkästään leksikaalisiin sanoihin perustuviin menetelmiin. Testauksessa käytettyä tiedonhakujärjestelmää (COATER-tiedonhakujärjestelmä) verrattiin SMART-tiedonhakujärjestelmään.

COATER määrittelee dokumenttien relevanttisuuden tutkimalla, kuinka yhdenmukaisia kyselyssä ilmaistut käsitteet ovat dokumentin tekstisisällön kanssa. Hakuoperaatiossa jokaiselle kyselyssä esiintyvälle käsitteelle määritellään sen esiintymiskonteksti ja kuinka dominantti kyseinen konteksti on kyseisessä dokumentissa. Hypoteesin mukaan tämä parantaa tarkkuutta, sillä useimmissa relevanteissa dokumenteissa dominantit kontekstit liittyvät kyselyssä ilmaistuihin käsitteisiin. Indeksointisanastona käytetään joukkoa WordNetin synonyymijoukkoja.

$Appeal_{1.1}$.....$court_{1.1}$.....$drug_{2.1}$.....$sentence_{1.1}$
.....$judge_{1.1}$.....$case_{1.1}$.....
$hearing_{1.1}$.....$cocaine_{2.1}$.....
: : : : : :
.....
.....
.....
: : : : : :
.....$importing_{3.1}$.....$narcotics_{2.2}$.....$illegal$
$substance_{2.1}$ .....$distribution_{3.1}$.....
...$network_{3.1}$.....$supply_{3.1}$.....$cocaine_{2.2}$.....
$cocaine_{2.2}$.....$heroin_{2.2}$
: : : : : :
.....
.....
.....
: : : : : :
$verdict_{1.2}$.....$judge_{1.2}$.....$prosecution_{1.2}$.....
.....$importation_{3.2}$.....$narcotics_{2.3}$.....
$cocaine_{2.3}$.....$sentence_{1.2}$.....
.....$suppliers_{3.2}$.....

Taulukko 10: Kolme leksikaalista kontekstia

Järjestelmiä testattiin pienessä dokumenttikokoelmassa. COATERin saavuttama tarkkuus oli suurempi kuin SMART-järjestelmän. COATERIN palauttamista dokumenteista 81 prosenttia oli relevantteja, kun taas SMARTin palauttamista dokumenteista vain 57 prosenttia oli relevantteja. Stairmandin mukaan käytännön sovellusten suhteen on kuitenkin huomioitava se seikka, että WordNetin kattavuus rajoittaa saavutettavaa tarkkuutta. Erisnimien puuttuminen on erityinen ongelma. WordNet ei sisällä myöskään kollokaatioinformaatiota. Kollokaatioinformaatiota voidaan kuitenkin saada dokumenttikokoelmaa varten räätälöidystä lisäresurssista.

Järjestelmän dokumenttirepresentaatiot muodostavat osittain merkityksen mukaan luokitellun korpuksen, ja niitä voidaan hyödyntää tuotettaessa käsiteryhmiä, jotka perustuvat kollokaatiosuhteisiin WordNetin käsitteiden välillä. Kollokaatiosuhde on kahden elementin välillä, jotka esiintyvät yhdessä tietyllä frekvenssillä. Sellaiset ryhmät tuotetaan tiettyjä dokumenttikokoelmia varten ja niitä voidaan hyödyntää tiedon haussa. Tiettyyn dokumenttikokoelmaan liittyvä resurssi on hyödyllinen kahdella tavalla. Kun jotain aihetta käsitellään dokumentissa, käytetään tyypillisesti aihepiiriin liittyviä termejä. Näiden termien välinen tarkka suhde voi kuitenkin olla vaikea määritellä, eikä spesifisiin käsitteiden välisiin suhteisiin perustuvista resursseista, kuten WordNetista ole välttämättä apua. Sellaisten suhteiden automaattinen derivointi dokumenttikokoelmasta voi potentiaalisesti johtaa dokumenttien aiheiden parempaan tunnistukseen. Toiseksi kyselyn rikastamisprosessilla voi olla resurssi leksikaalisen tiedon ulkoisiin lähteisiin, ja dokumenttikokoelmalle räätälöity resurssi voi mahdollistaa tarkemman ja laajemman kyselyn rikastamisen.

6.4 Yhteenveto

Tässä luvussa on tarkasteltu WordNetin hyödyntämistä tiedonhakusovelluksissa. WSD:n hyödyllisyyttä tiedonhaun kannalta koskevat tutkimukset ovat antaneet osin

ristiriitaisia tuloksia. Sandersonin [San94] ja Krovetzin ja Croftin [KC92] mukaan yksiselitteistäminen ei ole yleisesti ottaen hyödyllistä tiedonhakusovelluksissa. Gonzalon et. al [GVCC98] mukaan yksiselitteistäminen parantaa tiedonhakujärjestelmän suorituskykyä, ainakin jos virhemarginaali on suhteellisen pieni. WSD:n hyödyllisyys tiedonhakusovelluksissa saattaa riippua käytettävästä dokumenttikokoelmasta (hyödyllisempää heterogeenisessa dokumenttikokoelmassa kuin yhteen aihepiiriin liittyvässä) ja yksiselitteistettävien sanojen ominaisuuksista [KC92]. Mikäli käytettävissä on tarkka WSD-proseduuri, yksiselitteistäminen ei todennäköisesti ainakaan huononna hakutuloksia. WSD-algoritmit ovat kuitenkin laskennallisesti vaativia, mikä täytyy myös ottaa huomioon.

Semanttinen indeksointi, jossa indeksoidaan sanojen merkityksillä eikä leksikaalisilla merkkijonoilla, saattaa parantaa hakutuloksia ainakin tapauksissa, joissa sekä kyselyt että dokumentit ovat lyhyitä [SQ96]. Yhdistetyllä semanttisella ja sanamuotoon perustuvalla indeksoinnilla voi olla yleisemminkin potentiaalia tiedonhakujärjestelmissä [MM00]. WordNetin avulla muodostettujen leksikaalisten klusterien ja leksikaalisten ketjujen käyttö voi parantaa huomattavasti tiedonhakujärjestelmän tarkkuutta klassiseen sanamuotoon perustuvaan hakuun verrattuna [Sta97]. Käytännön sovellusten kannalta erisnimien puuttuminen WordNetista on kuitenkin ongelma.

7 WordNet ja dokumenttien luokittelu

Tässä luvussa tarkastellaan WordNetin hyödyntämistä dokumenttien luokittelussa. Luvussa 7.1 kuvataan yleisesti dokumenttien luokittelua koneoppimismenetelmillä. Luvuissa 7.2-7.4 käsitellään koneoppimiseen ja WordNet-tietokannan hyödyntämiseen perustuvia lähestymistapoja dokumenttien luokitteluun. Luvussa 7.2 tarkastellaan WordNet-informaation hyödyntämisen vaikutusta luokittelun tarkkuuteen erilaisissa dokumenttikokoelmissa [SM98]. Luvussa 7.3 käsitellään WordNetissa olevan tiedon liittämistä kahteen koneoppimislähestymistapaan, Rocchion ja Windrow-Hoffin algoritmeihin [dBRHA01]. Luvussa 7.4 tarkastellaan dokumenttien luokittelua itseorganisoituvan kartan ja WordNetin semanttisten verkkojen integraation avulla [WH02].

7.1 Dokumenttien luokittelu koneoppimismenetelmillä

Dokumenttien luokittelujärjestelmiä tarvitaan erilaisissa sovelluksissa, kuten sähköpostien ja uutisten suodattaminen, henkilökohtaiset informaatioagentit, tiedonhaku ja automaattinen indeksointi. Yleisesti ottaen luokittelun tarkoituksena on liittää dokumentit oikeaan aihepiiriin. Dokumenttien luokittelu perustuu tyypillisesti johonkin koneoppimisalgoritmiin.

Ohjattu koneoppiminen (supervised machine learning) voidaan määritellä seuraavasti: Käytettävissä on joukko luokkia C ja joukko oppimisesimerkkejä E, joista jokainen liitetään johonkin luokkaan. Järjestelmä käyttää oppimisesimerkkejä muodostamaan hypoteeseja, joita voidaan käyttää ennustamaan samantyyppisten, aikaisemmin näkemättömien esimerkkien luokat [Mit97, SM98]. Koneoppimisjärjestelmissä, jotka luokittelevat dokumentteja, E on korpuksen luokiteltujen dokumenttien joukko.

Ennenkuin joukko dokumentteja esitetään koneoppimisjärjestelmälle, jokainen dokumentti muunnetaan piirrevektoriksi. Tyypillisesti jokainen piirrevektorin elementti

edustaa korpuksessa esiintyvää sanaa. Piirrearvot voivat olla binäärisiä, osoittaen, että sana esiintyy tai ei esiinny dokumentissa. Ne voivat olla myös kokonaislukuja tai reaalilukuja, jotka kuvaavat sanan esiintymiskertoja tekstissä. Tästä esitystavasta käytetään nimitystä bag-of-words-esitystapa. Sitä käytetään monissa lähestymistavoissa dokumenttien luokitteluun [Lan95, Joa97, Col97]. Näissä lähestymistavoissa alkuperäistä tekstiä ei prosessoida muuten kuin käyttämällä hukkasanalistaa yleisimpien sanojen poistamiseen.

7.2 Dokumenttien luokittelu WordNetin hypernyymien avulla

Scott ja Matwin [SM98] kuvaavat dokumenttienluokittelumenetelmän, jossa hyödynnetään koneoppimista ja dokumenttien esittämistä WordNetin hypernyymien avulla. Luokittelujärjestelmän käyttämät diskriminaatiosäännöt tuotetaan Ripperin järjestelmän [Coh95] avulla. Säännöt tuotetaan sekä käyttäen *hypernyymitiheys* (hypernym density) -esitystapaa, että bag-of-words-esitystapaa, jolloin mukaan ei liitetä WordNetissa olevaa tietoa. Testit osoittivat, että joissakin vaikeissa luokittelutehtävissä hypernyymitiheyslähestymistapa johtaa selvästi parempiin tuloksiin.

Scott ja Matwin [SM98] tutkivat hypoteesia, jonka mukaan kielellisen tiedon liittäminen tekstin esitystapaan voi parantaa luokittelun tarkkuutta. Kielellinen tieto saadaan Ripperin luokittelijasta ja WordNetin synonyymi- ja hyponyymisuhteista. Tämän tiedon avulla tekstin esitystapa muunnetaan bag-of-words-esitystavasta hypernyymitiheysesitystavaksi. Scott ja Matwin esittävät tuloksia tutkimuksesta, jossa hypernyymitiheysesitystapaa sen erilaisissa yleisyyden asteissa verrataan bag-of-words-malliin.

7.2.1 Korpus ja luokittelutehtävät

Luokittelutehtävissä on käytetty kolmea eri korpusta: Reuters-21578, USENET ja Digital Tradition (DigiTrad). Aihepiirien otsikoita käytetään luokittelun perustana Reutersin korpuksessa. USENETIN luokittelun perustana käytetään uutisryhmien nimiä. DigiTrad on 6500 kansanlaulun kokoelma. Jokaiseen lauluun liittyy haun helpottamiseksi yksi tai useampia avainsanoja. Jotkut näistä avainsanoista sisältävät informaatiota laulun alkuperästä tai tyylistä, esimerkiksi *Irish* tai *British*. Toiset sisältävät informaatiota laulun aiheesta, esim. *murder* tai *marriage*. Jälkimmäistä tyyppiä olevia avainsanoja käytetään luokittelun perustana tässä tutkimuksessa.

Reutersin korpus koostuu faktuaalista informaatiota sisältävistä artikkeleista. Kirjoitustyyli on yksinkertainen ja sanasto on melko suppea. On havaittu, että Reutersin uutisotsikoilla on taipumus koostua sanoista, jotka esiintyvät usein varsinaisessa uutistekstissä. Tätä havaintoa on käytetty parantamaan luokittelun tarkkuutta [dBRHA01]. Digitrad ja USENET ovat esimerkkejä toisesta ääripäästä. DigiTradin teksteissä saatetaan käyttää metaforista tai epätavallista ja arkaaista kieltä. USENETIN kirjoittajat käyttävät usein vaihtelevaa terminologiaa, eksyvät asiasta tai käyttävät epätavallista kieltä. Kaikki tämä tekee aiheeseen perustuvan luokittelun DigiTradissa ja USENETissa vaikeammaksi kuin Reutersissa.

Taulukossa 11 kuvataan kuusi luokittelutehtävää, jotka on suoritettu jossakin näistä kolmesta dokumenttikokoelmasta. Kussakin tehtävässä dokumentit luokitellaan jompaankumpaan annetusta kahdesta luokasta, esimerkiksi 'livestock', tai 'gold'.

Käytetty koneoppimisalgoritmi on Ripperin luokittelija, jota Cohen on kehittänyt [Coh95, Coh96]. Kuten taulukosta 11 nähdään, virheprosentti Reutersin tapauksessa on alle 7 prosenttia sekä bag-of-words, että hypernyymitiheys lähestymistapoja käytettäessä. Muissa luokittelutehtävissä virheiden osuus vaihteli 19 prosentista 38 prosenttiin.

Tehtävän nimi	lähde	luokat	koko	esimerkkejä	sanoja	virhe %
REUTER1	Reuters-21578	livestock/gold	224	98/126	154	1.75
REUTER2	Reuters-21578	corn/wheat	313	130/183	173	3.87
SONG1	DigiTrag	murder/marriage	424	200/224	331	30.23
SONG2	DigiTrad	political/religion	432	194/238	241	32.64
USENET1	USENET	sos.history/ misc.taxes.moderated	249	79/170	166	19.92
USENET2	UNSENET	bionet.microbiology/ bionet.neuroscience	280	117/163	152	37.86

Taulukko 11: *Luokittelutehtävät. Koko tarkoittaa kussakin tehtävässä käytettävien dokumenttien kokonaismäärää. Esimerkkejä sarakkeessa on kummankin luokan esimerkkien määrät. Sanoja sarakkeessa on kunkin tehtävän dokumenttien keskipituus.*

7.2.2 Hypernyymitiheysesitystapa

Hypernyymitiheyden laskeminen korpuksesta tapahtuu kolmessa vaiheessa:

1) Brillin jäsennintä [Bri92] käytetään liittämään sanaluokkatunniste jokaiseen sanaan korpuksessa.

2) Korpuksessa esiintyvät substantiivit ja verbit etsitään WordNetista. Niistä synonyymijoukoista, joissa ne esiintyvät ja näiden synonyymijoukkojen hypernyymisynonyymijoukoista muodostetaan globaali lista. Harvoin esiintyvät synonyymijoukot poistetaan. Jäljellejäävät muodostavat piirrejoukon.

3) Jokaisen synonyymijoukon tiheys lasketaan kullekin koneoppimisesimerkille. Tuloksena on joukko numeerisia piirrevektoreita. Tiheys määritetään synonyymijoukon esiintymiskertoina tulosteessa jaettuna dokumentin sanojen määrällä.

Hypernyymitiheyden laskemisen aikana ei suoriteta yksiselitteistämistä. Kaikkia WordNetin palauttamia merkityksiä pidetään yhtä todennäköisinä, ja ne kaikki liitetään piirrejoukkoon. Oppimista auttaa se, että monet erilaiset, mutta synonyymiset ja hyponyymiset sanat liitetään samoihin synonyymijoukkoihin, mikä nostaa relevanteimpien synonyymijoukkojen tiheyksiä. Jos piirre saa suhteellisen alhaisen arvon, se tarkoittaa, että kyseisen synonyymijoukon merkityksellisyydestä dokumentille on vain vähän evidenssiä.

Tämän prosessin aikana voi tapahtua virheitä, jotka vaikuttavat piirrejoukkoon. Virhelähteitä ovat jäsennin, yksiselitteistämisen puute, sanojen puuttuminen WordNetista ja WordNetin semanttisen hierarkian mataluus joissakin aihepiireissä.

7.2.3 Testit ja tulokset

Uusi hypernyymitiheysesitystapa poikkeaa bag-of-words-esitystavasta siinä suhteessa, että normalisoidut tiheysvektorit korvaavat piirrevektorit ja hypernyymisynosyy-

mijoukot korvaavat sanamuodot. Testit suoritettiin seuraavia esitystapoja käyttäen: bag-of-words-esitystapa käyttäen verbejä ja substantiiveja, sekä normalisoidut tiheysvektorit substantiiveille ja verbeille.

Taulukossa 12 on esitetty kunkin luokittelutehtävän tulokset, kun on käytetty hypernyymintiheysesitystapaa h:n arvoilla 0-9 ja h = max. Taulukossa on esitetty kolmen virheasteen vertailu. Ne ovat bag-of-words, hypernyymin tiheys, kun h = max ja hypernyymin tiheys käyttäen parasta h:n arvoa.

Parametri h kontrolloi yleistyshierarkian korkeutta. Tätä parametria voidaan käyttää rajoittamaan kunkin sanan hypernyymihierarkiassa ylöspäin otettavien askelten määrää. Jos h = 0, lasketaan vain ne synonyymijoukot, jotka sisältävät korpuksessa esiintyvän sanan. h = max tarkoittaa, että kaikki hypernyymisynonyymijoukot haetaan, riippumatta siitä, kuinka korkealla hierarkiassa ne ovat.

Tehtävä	bag-of-words	Hypernyymintiheys			
		h	virheprosentti	h	virheprosentti
REUTER1	1.75	max	2.38	0	1.75
REUTER2	3.87	max	6.13	0	4.84
SONG1	30.23	max	22.04	9	*16.00*
SONG2	32.64	max	34.45	4	31.04
USENET1	19.92	max	*14.36*	9	*13.11*
USENET2	37.86	max	40.00	2	36.43

Taulukko 12: *Virheprosenttien vertailu 10-kertaisessa ristiinvalidoinnissa tutkimuksen kuudessa datajoukossa. Tilastollisesti merkittävät parannukset bag-of-words lähestymistapaan on kursivoitu*

Reutersin dokumenttikokoelmien tapauksessa ei ole havaittavissa parannusta bag-of-words lähestymistapaan verrattuna. Dokumenttikokoelmien SONG1 ja USENET1 tapauksissa virheprosentti pieneni selvästi. Dokumenttikokoelmien SONG2 ja USE-

NET2 tapauksissa hypernyymien käyttö taas tuotti bag-of-words lähestymistapaan verrattavat virheprosentit.

Dokumenttikokoelman SONG2 tapauksessa pääongelma näyttää olevan se, että luokat (political ja religion) ovat läheisessä semanttisessa suhteessa keskenään. Hypernyymintiheyssääntöä h = max käytettäessä löydettiin enimmäkseen hyvin abstrakteja synonyymijoukkoja, kuten {social group} ja {political unit}.

Dokumenttikokoelman USENET2 tapauksessa ongelmiin on ilmeisesti kaksi syytä. Luokat (microbiology ja neuroscience) ovat semanttisesti lähellä toisiaan. Toiseksi kirjoittajat käyttävät hyvin teknisiä termejä, joita ei ole käytetyssä WordNetin versiossa 1.5. Joitakin esimerkkejä puuttuvista sanoista ovat: *HIV*, *neurobiology* ja *retrovirus*.

7.2.4 Yhteenveto

Scottin ja Matwinin [SM98] mukaan hypernyymitiheysesitystapa voi toimia hyvin, jos dokumenteissa käytetty sanasto on laaja tai epätavallinen, tai eri kirjoittajat ovat käyttäneet erilaisia termejä samasta asiasta. Siitä ei todennäköisesti ole hyötyä, jos dokumenteissa käytetään suhteellisen suppeaa ja yhdenmukaista sanastoa. Kuitenkin, jos dokumenteissa käytetty sanasto on hyvin erikoistunutta, ongelmia saattaa syntyä siitä, että käytettyjä sanoja ei löydy WordNetista. Hypernyymitiheysesitystapa toimii todennäköisesti paremmin, jos dokumentit luokitellaan tarkasti määriteltyihin ja/tai semanttisesti kaukana toisistaan oleviin luokkiin. Jos luokat on määritelty laajasti ja/tai ne ovat semanttisesti lähellä toisiaan, hypernyymintiheysesitystapa ei välttämättä ole tarkempi kuin bag-of-words-esitystapa.

Tulevaisuudessa on tarkoitus käyttää myös muita WordNetissa esiintyviä suhteita, kuten meronymiaa. WSD:n avulla voidaan tuottaa tarkempia hypernyymintiheyspiirteitä. Muiden kielellisten resurssien, kuten Unified Medical Language System Metathesaurus käyttö voi parantaa luokittelijan suorituskykyä aihepiireissä, jotka ovat

semanttisesti lähellä toisiaan ja hyvin erikoistuneita [SM98].

7.3 WordNetin käyttö oppimisinformaation täydentämisessä dokumenttien luokittelussa

Buenaga et al. [dBRHA01] tarkastelevat WordNetissa olevan tiedon liittämistä kahteen koneoppimislähestymistapaan, Rocchion ja Windrow-Hoffin algoritmeihin. Todistaakseen hypoteesin, että leksikaalisten tietokantojen hyödyntäminen parantaa koneoppimiseen perustuvaa dokumenttienkategorisointijärjestelmää, Buenaga et al. suorittivat joukon kokeita Reuters-21578-tekstikoelmassa. Molempien hybridijärjestelmien arvioinnissa saavutetut tulokset osoittivat, että integroitu lähestymistapa, joka yhdistää oppimiskorpuksen ja leksikaalisen tietokannan on suorituskyvyltään parempi kuin pelkkä oppimiskorpuksen käyttö. Toiseksi integroitu lähestymistapa dokumenttien kategorisointiin voi auttaa dokumenttien luokittelua matalan frekvenssin kategorioihin, joille on vähän tai ei lainkaan oppimisdataa.

Rocchion algoritmia käytettäessä jokaiselle kategorialle luodaan eksplisiittinen profiili, eli prototyyppinen dokumentti. Kategorian profiili on vektori, jossa on yhtä monta ulottuvuutta, kuin dokumentteja kuvaavissa vektoreissa. Kunkin termin paino kuvaa kyseisen termin tärkeyttä kategorialle.

Rocchion algoritmi tuottaa uuden painovektorin wc_k olemassaolevasta painovektorista wc_k^0 ja oppimisdokumenttien kokoelmasta. Vektorin wc_k komponentti i lasketaan kaavalla:

$$wc_{ik} = \alpha we_{ik}^0 + \beta \frac{\sum_{l \in C_k} wd_{il}}{n_k} + \gamma \frac{\sum_{l \notin C_k} wd_{il}}{P - n_k}$$

missä wc_k^0 on kategorian k termin i alkuperäinen paino, wd_{il} on termin i paino oppimisdokumentille l, C_k on kategoriaan k liitettyjen dokumentti-indeksien joukko ja n_k

on näiden dokumenttien määrä [dBRHA01].

Widrow-Hoff-algoritmi aloittaa olemassaolevasta painovektorista wc_k^0 ja päivittää sitä joka kerran, kun uusi oppimisdokumentti lisätään. Vektorin wc_k^{l+1} komponentti i saadaan l:nnestä dokumentista ja l:nnestä vektorista kaavalla:

$$wc_{ik}^{l+1} = wc_{ik}^{l} + 2\eta(wd_l * wc_k^l - y\iota)wd_{il}$$

missä wc_{ik}^l on termin i paino kategorian k l:nnessä vektorissa, wd_l on termin painovektori dokumentille l, wc_k^l on kategorian k l:s vektori, $y\iota$ on 1, jos l:s dokumentti on liitetty kategoriaan k (ja 0 muissa tapauksissa) ja wd_{il} on termin i paino l:nnessä dokumentissa. Muuttuja η on oppimisnopeus, joka kontrolloi sitä, kuinka nopeasti painovektorin sallitaan muuttua ja kuinka paljon vaikutusta kullakin uudella dokumentilla on siihen. Tavallisesti käytetty η:n arvo on $1/4X_2$, missä X on oppimisdokumentteja edustavien vektorinormien maksimiarvo [dBRHA01].

7.3.1 Oppimisinformaation täydentäminen

WordNetista ja oppimiskokoelmasta saatavan informaation yhdistäminen suoritetaan asettamalla ensin alkupainot kategorioille. WordNetin hyödyntäminen perustuu oletamukseen, että kategorian nimi on hyvä ennuste sen esiintymiselle. Esimerkiksi sanan *barley* esiintyminen osoittaa, että uutisartikkeli tulisi luokitella kategoriaan BARLEY. Yleisempien kategorioiden, kuten EARN (*earnings*) esiintymisen ennustamisen tulee sen sijaan nojata semanttisesti itsenäisempien termien, kuten *dollar* tai *invest* esiintymiseen.

Tässä lähestymistavassa on keskitytty WordNetin synonyymisuhteisiin. Kunkin kategorian tapauksessa sen lähimmät synonyymijoukot valitaan, ja niihin kuuluvat termit lisätään termijoukkoon. Kandidaattisynonyymijoukkojen valinta suoritetaan manuaalisesti.

Termit, jotka eivät esiinny missään dokumentissa poistetaan. Jäljelle jääneille termeille lasketaan semanttinen läheisyys siihen kategoriaan, johon ne liittyvät seuraavien kriteerien mukaan:

Jos termi on kategoriaa edustavan ilmauksen suora synonyymi, kuten *peanut* on sanan *groutnut* synonyymi, semanttinen läheisyys termin ja kategorian välillä on 1.

Jos kategoriaa edustava ilmaus koostuu useista sanoista, minkä tahansa näiden sanojen synonyymin semanttinen arvo on 1/nc, missä nc on sanojen määrä ilmauksessa. Esimerkiksi termi *indicant* on synonyymi sanalle *index* ilmauksessa *industrial production index*, ja sen semanttinen arvo on 1/3. Jos kategorian ja termin välille voidaan määritellä useita arvoja, suurin niistä valitaan.

WordNetissa olevaa informaatiota on yhdistetty Rocchion ja Windrow-Hoffin algoritmeihin tuottamaan kategorioiden esitystapoja. Semanttiset läheisyysarvot otetaan kategorioiden alkupainoiksi. Nämä painot määritellään uudelleen oppimisdokumenttien avulla. Jotta alkupainot ja oppimisdokumenteista saadut painot pysyisivät samassa suuruusluokassa, kummallekin algortimille käytetään omaa menetelmää WordNet-informaation integrointiin.

Termien painot dokumenteissa ovat termien esiintymiskerrat kerrottuna termien painoilla oppimiskokoelmassa. Termien painot oppimiskokoelmassa lasketaan kaavalla:

$$\mathrm{tw_i} = log_2 \frac{P}{t\int_i}$$

missä $t\int_i$ on niiden oppimisdokumenttien määrä, joissa termi i esiintyy. P on oppimisdokumenttien kokonaismäärä. Rocchion algoritmin tapauksessa on huomioitu aikaisemmin tuotettu semanttisen läheisyyden arvo termin esiintymiskertoina kategoriassa. Tämä arvo kerrotaan termin painolla kokoelmassa. Windrow-Hoff-algoritmin tapauksessa termin painon lisääminen dokumenttiin normalisoidaan muuttujalla η.

7.3.2 Arviointia

Tarkkuuden laskentaa varten dokumentit on järjestetty kunkin katagorian tapauksessa sen mukaan, kuinka samanlaisia ne ovat kategorian kanssa. Kategoriat jaetaan kahteen ryhmään, kategoriat joissa on vähän oppimisesimerkkejä, ja kategoriat, joissa niitä on paljon. Tarkkuuden keskiarvot lasketaan jokaisella saantitasolla kummallekin kategoriajoukolle, ja kaikkien kategorioiden joukolle. Tällä tavoin jokaisella kategorialla on sama vaikutus lopputulokseen sen yleisyydestä riippumatta.

Koesarjan tulokset on esitetty taulukossa 13. Taulukossa nähdään tarkkuus 11 saantitasolla neljässä testatussa lähestymistavassa. Resursseja yhdistävillä lähestymistavoilla saavutettiin paljon parempia tuloksia kuin pelkkään koneoppimiseen perustuvilla lähestymistavoilla. WordNet-integraation avulla tarkkuus parani keskimäärin noin 20 pistettä molemmilla algoritmeilla. Kumpikaan näistä oppimisalgoritmeista ei osoittautunut merkittävästi toista paremmaksi.

Taulukossa 14 on esitetty keskitarkkuus kullekin testatulle lähestymistavalle. Keskitarkkuus on laskettu erikseen kategorioille, joille on vähemmän kuin 10 oppimisdokumenttia, ja kategorioille, joille oppimisdokumenttia on 10 tai enemmän.

7.4 Reutersin uutiskorpuksen itseorganisoiva luokittelu

Wermter ja Hung [WH02] tutkivat dokumenttien luokittelua *itseorganisoituvan kartan* ja WordNetin semanttisten verkkojen integraation avulla. Tämä neuroverkkomalli perustuu merkitys (significance)- vektoreihin. Tämän neuroverkon itseorganisoitumista ja hypernyymisuhteita yhdistävän lähestymistavan avulla saavutettiin hyviä luokittelutuloksia 100 000 artikkelin kokoelmassa. Nämä tulokset osoittavat, että tämä lähestymistapa voi sopia suuriin reaalimaailman tehtäviin.

Dokumenttien luokittelussa dokumentit ryhmitellään joukkoon ennalta määriteltyjä

Saanti	Koneoppiminen	Koneoppiminen	Koneoppiminen + WN	Koneoppiminen + WN
	Rocchio	WHoff	Rocchio	WHoff
0.0	0.567	0.565	0.733	0.703
0.1	0.478	0.484	0.703	0.659
0.2	0.423	0.427	0.661	0.610
0.3	0.362	0.375	0.601	0.555
0.4	0.315	0.331	0.573	0.530
0.5	0.270	0.279	0.556	0.511
0.6	0.224	0.225	0.503	0.469
0.7	0.175	0.179	0.416	0.436
0.8	0.147	0.149	0.359	0.412
0.9	0.119	0.122	0.296	0.351
1.0	0.109	0.111	0.201	0.289
Keski-arvo	0.290	0.295	0.509	0.502

Taulukko 13: *Tarkkuus 11 saantitasolla neljässä testatussa lähestymistavassa*

	< 10	$>= 10$	Kaikkiaan
Rocchio	0.276	0.297	0.290
Widrow-Hoff	0.278	0.305	0.295
Rocchio + WN	0.417	0.560	0.509
Widrow-Hoff + WN	0.482	0.514	0.502

Taulukko 14: *Tulokset kategorioille, joissa on vähän vs. paljon dokumentteja*

kategorioita. Perinteiset luokittelussa käytettävät neuroverkkomallit eivät voi esittää tuloksiaan helposti, ellei niihin liitetä lisämoduuleja tätä tarkoitusta varten [LSM91, Hon97]. Itseorganisoituvat käsitekartat voidaan kuitenkin esittää kaksiulotteisina, joten ne ovat helposti visualisoitavissa ja niitä on helppo tulkita. WordNetista eristetään sopivia suhteita edustamaan uutisartikkelien semanttista kenttää. Uutisartikkelien luokittelu suoritetaan käsitekartan ja WordNetin integraation avulla.

7.4.1 SOM

Luokittelujärjestelmässä hyödynnetään Kohosen [Koh82] esittämä neuroverkkomallia SOM (Selforganizing Memory Networks), joka perustuu ohjaamattomaan koneoppimiseen. Sen avulla voidaan pakata moniulotteinen datajoukko 2-ulotteiseen avaruuteen. SOM sijoittaa samankaltaisen datan topologisesti läheisille alueille käsitekartassa. Käyttäjät voivat sitten valita käsitekartan relevantit dokumenttiklusterit saadakseen relevantit dokumentit.

Vektorimalli on perustekniikka tekstidokumenttien muuntamiseen numeerisiksi vektoreiksi. Monet dokumenttien luokittelussa käytettävät neuroverkkomallit, mukaan lukien SOM soveltavat vektorimallia esiprosessointivaiheessa. Vektorimallissa dokumentit esitetään vektoreina moniulotteisessa avaruudessa, jossa jokainen ulottuvuus vastaa yhtä termiä. Yksittäiset dokumentit muodostetaan dokumenttia kuvaavien termien vektoreista. Dokumenteista eristetään yksittäiset sanat ja lasketaan niiden frekvenssit. Tuloksesta poistetaan hukkasanat ja päätteet. Tämän jälkeen lasketaan termien frekvenssit jokaisessa kokoelman dokumentissa ja valitaan indeksitermeiksi sanat, joiden frekvenssi ylittää tietyn kynnysarvon.

Käytettäessä WordNetin semanttisia suhteita yksi indeksitermi voi esittää termin monia synonyymeja, sisaruksia tai muita relevantteja termejä. WordNetia voidaan käyttää vähentämään ulottuvuuksia liittämällä sanat yleisempiin käsitteisiin.

Merkitysvektoreita käytetään esittämään termien tärkeyttä kussakin semanttisessa kategoriassa. Ennalta määriteltyjä aihepiirejä käytetään akseleina moniulotteisessa avaruudessa. Dokumenttia kuvaa n-ulotteinen vektori, jossa n on ennalta määriteltyjen aihepiirien määrä.

7.4.2 Itseorganisoiva luokittelu käyttäen WordNetia

Testeissä keskitytään kahdeksaan pää-aihepiiriin. Aihepiirit on kuvattu taulukossa 15.

no.	Aihepiiri	Kuvaus	Jakauma
1	C15	performance	149.358
2	C151	accounts/earnings	81.200
3	CCAT	corporate/industrial	372.097
4	E21	government finance	42.573
5	ECAT	economics	116.205
6	GCAT	government/social	232.031
7	GCRIM	crime, law enforcement	32.036
8	GDIP	international relations	37.630

Taulukko 15: *Valitut aihepiirit ja niiden jakauma koko korpuksessa*

Koska pelkästään substantiiveilla ja verbeillä on hypernyymisuhteita WordNetissa, ja substantiivit ja verbit sisältävät riittävästi informaatiota dokumentissa esiintyvistä käsitteistä, testeissä käytetään vain WordNetissa esiintyviä substantiiveja ja verbejä.

Jokaiselle uutisartikkelille määritellään aihepiiriluokitus. Tämä luokitus määritellään syötevektorin merkittävimpien aihepiirien perusteella. Sitten syötevektorit normalisoidaan. Oppimisprosessin jälkeen karttayksikköön liitetään luokitus sen mukaan, mitä luokituksia karttayksikköön on liitetty eniten. Esimerkiksi, jos 3 BCAT-uutisartikkelia ja 10 CCAT-uutisartikkelia liitetään karttayksikköön 1, silloin karttayksikkö 1 luo-

kitellaan CCAT:iin kuuluvaksi. Saman käsitteen sisältävät uutisartikkelit ryhmitellään samaan luokkaan, ja jokaisella ryhmän jäsenellä, eli jokaisella artikkelilla on yhä oma spesifi merkityksensä. Kaksitasoista hypernyymiä käytetään korvaamaan jokainen artikkelissa esiintyvä sana sen hypernyymitermillä, jotta saataisiin alkuperäistä sanaa yleisempi käsite. Polyseemiset ja synonyymiset termit voivat olla edustettuina useissa synonyymijoukoissa, ja jokainen näistä synonyymijoukoista voi sijaita eri hyponyymihierarkiassa. On vaikeaa päätellä, mihin kategoriaan kuuluu dokumentti, joka sisältää useita monimerkityksellisiä sanoja. Termien korrekti merkitys päätellään vertaamalla kunkin sanan selityksen merkitystä Reutersin semanttiseen termi-aihepiiri-tietokantaan. Esimerkiksi ensimmäinen uutisartikkeli on liitetty aluksi aihepiiriin ECAT. Tämän artikkelin otsikon ensimmäinen termi on *recovery*, jolla on WordNetissa 3 substantiivimerkitystä, eikä yhtään verbimerkitystä. Termin yhteisesiintymisten määrä kussakin sanan selityksessä ja esiliitetyssä termi-aihepiiri-tietokannassa lasketaan. Termien merkittävyyden keskiarvo lasketaan jakamalla saatu lukumäärä termien kokonaismäärällä kussakin sanan selityksessä. Lopuksi jokainen termi korvataan sen tason 2 hypernyymillä. Tämän lähestymistavan avulla onnistuttiin vähentämään erilaisten sanojen kokonaismäärää 83.15 prosenttia kokonaisten tekstien oppimisjoukossa ja 72.84 prosenttia otsikoiden oppimisjoukossa. Lukumäärät on esitetty taulukossa 16. Tämä lähestymistapa voi tarjota myös helpon tavan löytää suhteellisen oikea merkitys moniselitteiselle sanalle.

Oppimisjoukko	Ilman WordNetia	WordNetin avulla	Väheneminen prosentteina
Otsikot	10185	2766	72.84
Koko teksti	22848	3851	83.15

Taulukko 16: *Erilaisten sanojen kokonaismäärä oppimisjoukossa WordNetin avulla ja ilman*

7.4.3 Tulokset

Wermter ja Hung [WH02] esittävät kuuden testin tulokset. Neljässä ensimmäisessä testissä käytetään vain SOM:in perustuvaa luokittelua. Ensimmäisessä testissä käytetään 100 000 uutisotsikkoa opetusaineistona ja toista 100 000 uutisotsikkoa testaukseen. Toinen testi on muuten sama, mutta otsikoiden sijasta käytetään koko tekstiä. Kolmannessa testissä käytetään kokonaisia tekstejä opetusaineistona ja niiden otsikoita testaukseen, neljännessä testissä päinvastoin. Viidenessä ja kuudennessa testissä hyödynnetään SOM:n ja WordNetin integraatiota. Viidennessä testissä oppimisjoukkona on käytetty kokonaisia artikkeleja ja kuudennessa otsikoita.

Neljän ensimmäisen testin tulokset on esitetty taulukoissa 17 - 20. Vaikka kokonaiset artikkelit sisältävät enemmän informaatiota, kuin otsikot, tuloksissa ei ole suurta eroa.

Merkitysvektorit perustuvat sanan frekvenssiin eri aihepiireissä. Menetelmän 1 tapauksessa kuhunkin aihepiiriin liitettyjen dokumenttien määrä voi vaikuttaa tulokseen. Menetelmän 2 tapauksessa jakaumien mahdollista vinoutta on pyritty lieventämään.

Menetelmä	Oppimisjoukko	Testijoukko
1	88.85	87.55
2	91.07	89.03

Taulukko 17: *Tarkkuus 100 000 uutisotsikon tapauksessa oppimis- jatestijoukossa*

Menetelmä	Oppimisjoukko	Testijoukko
1	85.70	85.96
2	92.77	92.01

Taulukko 18: *Tarkkuus 100 000 kokonaisen uutisartikkelin tapauksessa oppimis- ja testijoukossa*

Menetelmä	Oppimisjoukko	Testijoukko
1	85.70	80.81
2	92.77	80.18

Taulukko 19: *Tarkkuus, kun kokonaisia artikkeleita on käytetty oppimisjoukkona ja otsikoita testijoukkona*

Menetelmä	Oppimisjoukko	Testijoukko
1	88.85	84.11
2	91.07	89.95

Taulukko 20: *Tarkkuus, kun otsikoita on käytetty oppimisjoukkona ja kokonaisia artikkeleita testijoukkona*

SOM:n ja WordNetin integraation avulla saavutetut tulokset osoittautuivat merkittäviksi. Erilaisten sanojen kokonaismäärä väheni 10 185:stä 2766:teen otsikoiden tapauksessa, ja 22 848:sta 3851:teen kokonaisten artikkelien tapauksessa. Toiseksi WordNetin käyttö neuroverkkotekniikan apuna paransi uutisartikkelien luokittelun tarkkuutta. Taulukoissa 21 ja 22 nähdään SOM-tekniikan avulla saavutettu tarkkuus WordNetin avulla ja ilman.

Menetelmä	SOM	SOM ja WordNet
1	85.70	92.21
2	92.77	98.95

Taulukko 21: *Tarkkuus WordNetin toisen tason hypernyymien avulla ja ilman, kun oppimisjoukkona on käytetty kokonaisia artikkeleja*

Wermterin ja Hungin mukaan tämä osoittaa, että WordNetin hypernyymisuhde sopii käytettäväksi dokumenttien luokittelussa. Tätä suhdetta käyttämällä tekstin luokittelijan suorituskyky parani. Tilastollisia neuroverkkomenetelmiä ja semanttisia sym-

Menetelmä	SOM	SOM ja WordNet
1	88.85	89.94
2	91.07	90.65

Taulukko 22: *Tarkkuus WordNetin toisen tason hypernyymien avulla ja ilman, kun oppimisjoukkona on käytetty otsikoita*

bolisia suhteita yhdistävä neuroverkkokoneoppimistekniikka pystyy luokittelemaan tekstidokumentit yli 90-100 prosentin tarkkuudella oikeaan aihepiiriin.

7.5 Yhteenveto

Tässä luvussa on tarkasteltu WordNet-informaation hyödyntämistä dokumenttien luokittelussa. Oppimiskorpuksesta saatavaa tietoa ja WordNet-informaatiota hyödyntävä menetelmä dokumenttien luokitteluun voi olla hyödyllinen erityisesti, jos dokumenteissa käytettävä sanasto on suhteellisen monipuolinen ja kirjoitustyyli ei ole kovin yhdenmukainen, tai kategorialle on vähän tai ei lainkaan opetusaineistoa.

WordNetin kattavuus saattaa olla ongelma, jos dokumenttien sanasto on hyvin erikoistunutta. Lisäresursseista voi kuitenkin olla apua tässä asiassa.

Luvussa 7.3 tarkasteltiin WordNet-informaation integrointia kahteen koneoppimisalgoritmiin. Tulosten perusteella WordNet-informaation integrointi parantaa selvästi dokumenttien luokittelun tarkkuutta.

Käsitekarttaan perustuvassa dokumenttien luokittelussa sanat voidaan WordNetin hypernyymisuhteiden avulla liittää yleisempiin käsitteisiin ja näin vähentää ulottuvuuksia. WordNet-informaation hyödyntäminen lisää myös tämän menetelmän tapauksessa luokittelun tarkkuutta.

8 Johtopäätökset

Tässä tutkielmassa on käsitelty WordNet-sanatietokannan hyödyntämistä NLP-sovelluksissa. Yleisesti ottaen sen kattavuus ei useimmissa aihepiireissä ole ongelma. Noin 90 prosenttia teksteissä esiintyvistä substantiiveista löytyy WordNetista. Erisnimien puuttuminen WordNetista aiheuttaa kuitenkin ongelmia joissakin sovelluksissa. WordNetin rakenteen suhteen ongelmia aiheutuu mm. sanaluokkien välisten semanttisten suhteiden puuttumisesta, mikä estää käsittelemästä muita sanoja kuin substantiiveja joissakin sovelluksissa. WordNetin merkityserottelujen hienojakoisuus saattaa muodostua ongelmaksi joissakin yksiselitteistämismenetelmissä. Käsitteiden välinen etäisyys WordNetin hierarkiassa ei aina ole suorassa suhteessa käsitteiden semanttiseen läheisyyteen. Toisiinsa semanttisessa suhteessa olevien käsitteiden väliltä voi myös puuttua linkkejä.

Luvussa 3 tutustuttiin erilaisiin disambiguaatiotekniikoihin. Yleisesti ottaen WordNetin soveltuvuudesta yksiselitteistämiseen voi sanoa, että saavutettu tarkkuus ei ole yhtä hyvä kuin parhailla koneoppimiseen perustuvilla tekniikoilla. Niillä on kuitenkin käyttöä joissakin interaktiivisissa NLP-sovelluksissa, joissa ei ole käytettävissä manuaalisesti luokiteltua oppimiskorpusta, ja/tai koneoppimisprosessiin ei ole aikaa. Joissakin sekä WordNetia että koneoppimista hyödyntävissä menetelmissä käytetään manuaalisesti luokittelematonta korpusta [LCM98, MM98].

Substantiiviryhmien yksiselitteistämismenetelmissä WordNetin merkityserottelujen hienojakoisuus saattaa olla ongelma. Tämä ongelma voidaan ratkaista mahdollistamalla yksiselitteistäminen suhteessa korkeamman tason WordNet-kategorioihin [Res95]. Käsitteiden semanttiseen läheisyyteen ja niiden väliseen etäisyyteen WordNetissa liittyvä mahdollinen epäjohdonmukaisuus on pyritty ratkaisemaan Resnikin [Res95] esittämässä yksiselittesitämismenetelmässä pitämällä käsitteiden välisen samankaltaisuuden kriteerinä niiden yhteisen yläkäsitteen informatiivisuutta niitä yhdistävän polun

pituuden sijasta.

Luvussa 5 käsiteltiin WordNetin hyödyntämistä leksikaalisten ketjujen muodostamisessa. Barzilayn ja Elhadadin [BE97] esittämässä menetelmässä leksikaalisten ketjujen muodostamiseen hyödynnetään myös WordNetin substantiivitietokantaan sisältyviä substantiiviyhdistelmiä. Heidän mukaansa substantiiviyhdistelmien tunnistaminen on kahdella tavalla hyödyllistä: Sen avulla voidaan tunnistaa tärkeitä aihpiiriin liittyviä käsitteitä ja substantiiviyhdistelmien modifioijina esiintyvät sanat voidaan eliminoida, niin että niitä ei pidetä mahdollisina ketjun jäseninä.

Käsitteiden välinen etäisyys WordNetin hierarkiassa ei aina ole suorassa suhteessa käsitteiden semanttiseen läheisyyteen. Toisiinsa semanttisessa suhteessa olevien käsitteiden väliltä voi myös puuttua linkkeja. Tämä voi aiheuttaa virheitä leksikaalisten ketjujen muodostamisessa.

Luvussa 6 käsiteltiin WordNet tietokannan hyödyntämistä tiedonhakusovelluksissa. Niissä WordNetia on hyödynnetty kolmella tavalla; Dokumenteissa ja kyselyissä esiintyvien sanojen yksiselitteistämisessä, kyselyjen rikastamisessa ja semanttisessa indeksoinnissa. Käsitykset yksiselitteistämisen hyödyllisyydestä tiedonhaun kannalta ovat osittain ristiriitaisia. Mikäli käytettävissä on tarkka WSD-proseduuri, yksiselitteistäminen ei todennäköisesti ainakaan huononna hakutuloksia. Kyselyjen rikastamisesta voi olla hyötyä, jos kyselyt ovat lyhyitä ja epätäydellisiä. Pelkkään semanttiseen indeksointiin perustuva haku ei yleisesti ottaen anna parempia tuloksia kuin sanamuotoon perustuvaan indeksointiin perustuva haku. Yhdistetty semanttinen ja sanamuotoon perustuva indeksointi sen sijaan parantaa tuloksia ainakin vähän. WordNetin avulla muodostettujen leksikaalisten klusterien ja leksikaalisten ketjujen käyttö voi parantaa huomattavasti tiedonhakujärjestelmän tarkkuutta klassiseen sanamuotoon perustuvaan hakuun verrattuna. Kattavuuden suhteen erisnimien puuttuminen WordNetista voi kuitenkin olla ongelma.

Luvussa 7 käsiteltiin WordNetin hyödyntämistä dokumenttien luokittelussa. Oppimiskorpuksesta saatavaa tietoa ja WordNet-informaatiota hyödyntävä menetelmä dokumenttien luokitteluun voi olla hyödyllinen erityisesti, jos dokumenteissa käytettävä sanasto on suhteellisen monipuolinen ja kirjoitustyyli ei ole kovin yhdenmukainen. WordNetin kattavuus saattaa toisaalta olla ongelma, jos dokumenttien sanasto on hyvin erikoistunutta. Lisäresursseista voi olla apua tässä asiassa. WordNet-informaation hyödyntäminen luokittelussa voi myös olla hyödyllistä tapauksissa, joissa katagorialle on vähän tai ei lainkaan oppimisdataa.

Yleisesti ottaen WordNetista/leksikaalisesta tietokannasta on ilmeisesti eniten hyötyä leksikaalisten ketjujen muodostamisessa ja dokumenttien luokittelussa. Yksiselitteistämisessä ja tiedon haussa siitä voi olla hyötyä joissakin yksittäisissä menetelmissä.

Lähteet

AHK98 Reem Al-Halimi and Rick Kazman. Temporal indexing through lexical chaining. In Christiane Fellbaum, editor, *WordNet: an Electronic Lexical Database*, pages 333–353. Cambridge, MA: The MIT Press, 1998. `http://citeseer.nj.nec.com/73237.html`.

AR95 Eneko Agirre and German Rigau. A proposal for word sense disambiguation using conceptual distance. In *Proceedings of the 1st International Conference on Recent Advances in Natural Language Processing*, 1995. `http://xxx.lanl.gov/abs/cmp-lg/9510003`.

AR96 Eneko Agirre and German Rigau. Word sense disambiguation using conceptual density. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 16–22, 1996. `http://xxx.lanl.gov/abs/cmp-lg/9606007`.

Aun02 Lili Aunimo. Tekstifragmenttien välisen semanttisen samanlaisuuden tunnistaminen. Master's thesis, Helsingin yliopisto, 2002. `http://ethesis.helsinki.fi/julkaisut/hum/yleis/pg/aunimo/`.

BE97 Regina Barzilay and Michael Elhadad. Using lexical chains for text summarization. In *Proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS'97), ACL*, pages 10–17, Madrid, Spain, 1997. `http://citeseer.nj.nec.com/barzilay97using.html`.

Bri92 Eric Brill. A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, pages 152–155, Trento, IT, 1992. ACL. `http://citeseer.nj.nec.com/84525.html`.

Coh95 William W. Cohen. Fast effective rule induction. In *Proc. 12th International Conference on Machine Learning*, pages 115–123. Morgan Kaufmann, 1995. http://citeseer.nj.nec.com/cohen95fast.html.

Coh96 William W. Cohen. Learning trees and rules with set-valued features. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence and the Eighth Innovative Applications of Artificial Intelligence Conference*, pages 709–716, Menlo Park, August 4–8 1996. AAAI Press / MIT Press. http://citeseer.nj.nec.com/56507.html.

Col97 Sahami Coller. Hierarchically classifying documents using very few words. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML-97)*, pages 170–176, 1997. http://citeseer.nj.nec.com/koller97hierarchically.html.

dBRHA01 Manuel de Buenaga Rodríguez, José M. Gómez Hidalgo, and Belén Díaz Agudo. Using wordnet to complement training information in text categorization. In *Recent Advances in Natural Language Processing II: Selected Papers from RANLP'97*, volume 189, pages 353–364, 2001. http://arxiv.org/abs/cmp-lg/9709007.

Gre99a Stephen J. Green. Building hypertext links by computing semantic similarity. In *IEEE Transactions on Knowledge and Data Engineering*, pages 713–730, sept/oct 1999. http://ftp.cs.toronto.edu/pub/gh/Green-99.pdf.

Gre99b Stephen J. Green. Lexical semantics and automatic hypertext construction. *ACM Computing Surveys*, 1999. http://www.cs.brown.edu/memex/ACM_HypertextTestbed/papers/48.html.

GVCC98 Julio Gonzalo, Felisa Verdejo, Irina Chugur, and Juan Cigarran. Indexing

with WordNet synsets can improve text retrieval. In *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, pages 38–44, 1998. http://arxiv.org/ps/cmp-lg/9808002.

Hea94 Marti Hearst. Multi-paragraph segmentation of expository text. In *32nd. Annual Meeting of the Computational Linguistics*, pages 9–16, New Mexico State University, Las Cruces, New Mexico, 1994. http://citeseer.nj.nec.com/hearst94multiparagraph.html.

HH76 Michael Halliday and Ruqaiya Hasan. *Cohesion in English*, volume 9 of *English language series*. Longman, London, 1976.

Hon97 Timo Honkela. *Self-Organizing Maps in Natural Language Processing*. PhD thesis, Helsinki University of Technology, Espoo, Finland, 1997. http://www.cis.hut.fi/~tho/thesis/.

HSO98 Graeme Hirst and David St-Onge. Lexical chains as representations of context for the detection and correction of malapropisms. In Christiane Fellbaum, editor, *WordNet: an Electronic Lexical Database*, pages 305–333. Cambridge, MA: The MIT Press, 1998. http://citeseer.nj.nec.com/hirst97lexical.html.

JM91 Graeme Hirst Jane Morris. Lexical cohesion competed by thesaural relations as the indicator of the structure of text. *Computational Linguistics*, 17(1):21–48, 1991. http://www.cs.mu.oz.au/acl/J/J91/J91-1002.pdf.

Joa97 Thorsten Joachims. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. In Douglas H. Fisher, editor, *Proceedings of ICML-97, 14th International Conference on Machine Lear-*

ning, pages 143–151, Nashville, US, 1997. Morgan Kaufmann Publishers, San Francisco, US. `http://www-ai.cs.uni-dortmund.de/DOKUMENTE/joachims_97a.ps.gz`.

KC92 Robert Krovetz and W. Bruce Croft. Lexical ambiguity and information retrieval. *ACM Transactions on Information Systems*, 10(2):115–141, April 1992.

Koh82 Teuvo Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, (43):59–69, 1982.

Lan95 Ken Lang. Newsweeder: Learning to filter news. In *Proceedings of the 12th International Conference on Machine Learning*, pages 331–339, 1995. `http://citeseer.nj.nec.com/lang95newsweeder.html`.

LCM98 Claudia Leacock, Martin Chodorow, and George A. Miller. Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics*, 24(1):147–165, 1998. `http://acl.ldc.upenn.edu/J/J98/J98-1006.pdf`.

Lev93 Beth Levin. *English Verb Classes and Alternations*. The University of Chicago Press, 1993.

LSM91 Xia Lin, Dagobert Soergel, and Gary Marchionini. A self-organizing semantic map for information retrieval. In *Proceedings of the Fourteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Semantic Models, pages 262–269, 1991. `http://www.acm.org/pubs/articles/proceedings/ir/122860/p262-lin/p262-lin.pdf`.

MBF+93 George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. Indroduction to WordNet: An on-line lexical

database. *International Journal of Lexicography*, 3(4):235–244, 1993. http://engr.smu.edu/~rada/wnb/#M.

Mil98a George Miller. Nouns in WordNet. In Christiane Fellbaum, editor, *WordNet: an Electronic Lexical Database*, pages 23–47. Cambridge, MA: The MIT Press, 1998.

Mil98b Katherine Miller. Modifiers in WordNet. In Christiane Fellbaum, editor, *WordNet: an Electronic Lexical Database*, pages 47–69. Cambridge, MA: The MIT Press, 1998.

Mit97 Tom Mitchell. *Machine learning.* New York (NY) : McGraw-Hill, 1997.

MM98 Rada Mihalcea and Dan I. Moldovan. Word sense disambiguation based on semantic density. In Sanda Harabagiu, editor, *Proceedings of the Workshop on Usage of WordNet in Natural Language Processing Systems, COLING-ACL-98*, pages 16–22. Association for Computational Linguistics, Somerset, New Jersey, 1998. http://citeseer.nj.nec.com/mihalcea98word.html.

MM00 Rada Mihalcea and Dan Moldovan. Semantic indexing using WordNet senses. In *Proceedings of ACL Workshop on IR and NLP, October 2000*, 2000. http://citeseer.nj.nec.com/417656.html.

MS88 Marc Moens and Mark Steedman. Temporal ontology and temporal reference. *Computational Linguistics*, pages 15–28, 1988. http://acl.ldc.upenn.edu/J/J88/J88-2003.pdf.

Pus95 James Pustejovsky. *The Generative Lexicon.* Cambridge (MA), Mit Press, 1995.

Res95 Philip Resnik. Disambiguating noun groupings with respect to WordNet senses. In David Yarovsky and Kenneth Church, editors, *Proceedings of the Third Workshop on Very Large Corpora*, pages 54–68, Somerset, New Jersey, 1995. Association for Computational Linguistics. `http://acl.ldc.upenn.edu/W/W95/W95-0105.pdf`.

RM00 Dan Moldovan Rada Mihalcea. An iterative approach to word sense disambiguation. In *Proceedings of The FLorida Artificial Intelligence Research Society (FLAIRS-2000)*, pages 219–223, Orlando, FL, may 2000. `http://www.cs.ualberta.ca/~tszhu/resource/paper/wordsense/`.

RMBB89 R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1):17–30, 1989.

RS97 Ray Richardson and Alan Smeaton. Using WordNet in a knowledge-based approach to information retrieval. Technical report, School of Computer Applications, Dublin City University, 1997. `http://citeseer.nj.nec.com/richardson95using.html`.

San94 Mark Sanderson. Word sense disambiguation and information retrieval. In *Proceedings of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval*, pages 49–57, Dublin, IE, 1994. `http://citeseer.nj.nec.com/sanderson96word.html`.

Sie98 Eric V. Siegel. Disambiguating verbs with the wordnet category of direct object. In Sanda Harabagiu, editor, *Proceedings of the Workshop on Usage of WordNet in Natural Language Processing Systems, COLING-ACL-98*, pages 9–15. Association for Computational Linguistics, Somerset, New Jersey, 1998. `http://acl.ldc.upenn.edu/W/W98/W98-0702.pdf`.

SM98 Sam Scott and Stan Matwin. Text classification using WordNet hypernyms. In Sanda Harabagiu, editor, *Proceedings of the Workshop on Usage of WordNet in Natural Language Processing Systems, COLING-ACL-98*, pages 38–44. Association for Computational Linguistics, Somerset, New Jersey, 1998. `http://citeseer.nj.nec.com/sam98text.html`.

SP95 Hinrich Schutze and Jan Pedersen. Information retrieval based on word senses. In *Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, Las Vegas NV, 1995.

SQ96 Alan Smeaton and Ian Quigley. Experiments on using semantic distances between words in image caption retrieval. In *Proceedings of the 19th International Conference on Research and Development in Information Retrieval*, pages 174–180, 1996. `http://citeseer.nj.nec.com/smeaton96experiments.html`.

Sta97 Mark Stairmand. Textual context analysis for information retrieval. In *SIGIR '97: Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 140–147, Philadelphia, PA, USA, July 27-31 1997. ACM. `http://delivery.acm.org/10.1145/260000/258552/p140-stairmand.pdf?key1=258552&key2=8759762601&coll=portal&dl=ACM&CFID=11948300&CFTOKEN=75281833`.

Sus93 Michael Sussna. Word sense disambiguation for free-text indexing using a massive semantic network. In *Proceedings of the 2nd International Conference on Information and Knowledge Management*, pages 67–74. ACM Press, 1993.

Ven67 Zeno Vendler. Verbs and times. In *Linguistics in Philosophy*. Cornell University Press, 1967.

Voo94 Ellen M. Voorhees. Query expansion using lexical-semantic relations. In *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 61–69, 1994. `http://www.acm.org/pubs/articles/proceedings/ir/188490/p61-voorhees/p61-voorhees.pdf`.

WH02 Stefan Wermter and Chihli Hung. Selforganizing classification on the reuters news corpus. In *Proceedings of COLING-02, the 19th International Conference on Computational Linguistics*, Taipei, TW, 2002. `http://www.his.sunderland.ac.uk/ps/coling-232.pdf`.

WOB98 Janyce Wiebe, Tom O'Hara, and Rebecca Bruce. Constructing bayesian networks from WordNet for word-sense disambiguation: Representational and processing issues. In Sanda Harabagiu, editor, *Use of WordNet in Natural Language Processing Systems: Proceedings of the Conference*, pages 23–30. Association for Computational Linguistics, Somerset, New Jersey, 1998. `http://citeseer.nj.nec.com/wiebe98constructing.html`.

Yar95 David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 189–196, Cambridge, MA, 1995. `http://www.cs.jhu.edu/~yarowsky/acl95.ps`.