ARTIKKELI

Pertti Väyrynen, Tapio Seppänen,
*Kai Noponen, & Ilkka Juuso**

Kielellisen tiedon hyödyllisyydestä kieliteknologian eri sovellusalueilla

Pertti Väyrynen et. al: Kielellisen tiedon hyödyllisyydestä kieliteknologian eri sovellusalueilla [On the usefulness of linguistic knowledge in different area of application in Language Tecnology] Informaatiotutkimus 21 (3), 59-66.

In this article, the usefulness of linguistic knowledge in different areas of application in language technology is discussed focusing on one area of application, namely, Information Retrieval.

Address: Pertti Väyrynen,, MediaTeam Oulu Group, University of Oulu, Erkki Koiso-Kanttilan katu 3, FIN-90014 University of Oulu, Finland, e-mail: pav@ee.oulu.fi.

1. Johdanto

Kieliteknologian sovellusalueilla kuten tiedonhaussa *lingvistisellä* tiedolla eli tiedolla kielestä ja puheesta katsotaan usein olevan tärkeä rooli mitä tulee hakutulosten parantamiseen tulevaisuudessa. *Luonnollisen kielen prosessointia* (Natural Language Processing) kokoelmana erilaisia kielen automaattisia käsittelymenetelmiä kielen eri rakennetasoilla (esim. syntaktinen analyysi, sanan sanaluokan tunnistus, etc.) on jo hyödynnetty jossain määrin tiedonhaussa suurten tekstimassojen automaattisessa prosessoinnissa. Lupaavista tuloksista huolimatta monet tutkijat uskovat kuitenkin, että hakutuloksia parannetaan (lyhyellä tähtäimellä) enemmän kehittämällä edelleen tilastollisia menetelmiä kuin soveltamalla luonnollisen kielen prosessointiin perustuvia menetelmiä. Saadut tutkimustulokset näyttäisivät Harman et al.:n (1996) mukaan vahvistavan tämän, vaikka poikkeuksiakin löytyy.

Tilastollisten menetelmien suosiminen johtuu eri syistä, joista kaikki eivät liity kielellisen tiedon hyödynnettävyyteen kieliteknologiassa, vaan esim. tendenssiin hyödyntää jo olemassa olevaa teknologiaa sellaisenaan, mikä johtaa usein siihen, että potentiaalisesti hyödyllinen kielellinen tieto usein vain lisätään systeemiin sellaisenaan ja toivotaan sen parantavan sovellusten suorituskykyä. Eräs keskeinen ongelma kielellisen tiedon hyödyntämisessä kieliteknologian eri sovellusalueilla onkin juuri lingvistisen tiedon integrointi valtavirtaa edustaviin tilastollisiin menetelmiin. Joidenkin tutkijoiden mielestä, esim. Huckvale (1996, 11), vain sellainen lingvistinen tieto parantaa sovellusten suorituskykyä, jota voidaan implementoida tilastollisilla menetelmillä.

Tämän artikkelin tarkoitus on pohtia potentiaalisesti hyödyllisen lingvistisen tiedon laajamittaisempaa hyödyntämistä kieliteknologian joillakin keskeisillä sovellusalueilla ja miten sen avulla voitaisiin parantaa sovellusten suorituskykyä. Esimerkkisovellusalueena tarkastellaan lingvistisen tiedon käyttöä tiedonhaussa.

*Kirjoittajat työskentelevät Oulun yliopistossa MediaTeam-tutkimusryhmässä sen Language and Audiology Team:ssä. Lisätietoja osoitteesta http://www.mediateam.oulu.fi?/lang=fi

2. Lingvistisen tiedon käytöstä kieliteknologiassa tällä hetkellä

2.1. Kielen rakennetasomalli

Lingvistiikassa kielellistä tietoa kuvataan yleensä kielen *rakennetasomallilla*, jossa kielellisiä tasoja voidaan erottaa kuudesta kahdeksaan kappaletta riippuen niiden käyttötarkoituksesta (Crystal 1997, 83). Esimerkiksi Schmandt (1994, 9) erottaa seuraavat kahdeksan lingvististä tasoa:

Diskurssitaso
Pragmaattinen taso
Semanttinen taso
Syntaktinen taso
Leksikaalinen taso
Foneemitaso
Artikulatorinen taso
Akustinen taso

Kuva 1. Kielen rakennetasot

Kuten Schmandt (1994) itsekin huomauttaa, ylläkuvattu malli on lähinnä analyyttinen eikä sellaisenaan kata kaikkia potentiaalisesti hyödyllisiä kielellisiä ilmiöitä.

Siirryttäessä kielen alemmalta rakennetasolta ylemmälle tasolle analyysiyksikön koko kasvaa ja itse analyysi tulee monimutkaisemmaksi ja vaikeammaksi toteuttaa. Mitä suurempi analyysiyksikkö (esim. morfeemi, sana, virke, kappale tai koko tekstidokumentti), sitä vaikeammin mallinnettavia usein abstraktit kielelliset ilmiöt ovat luonteeltaan ja sitä enemmän niihin sisältyy valinnaisuutta sekä variaatiota yleensäkin. (Liddy 1998, 2.)

Sinällään tilastollisesti harvinaisemmat abstraktit kielelliset ilmiöt, jotka tyypillisesti esiintyvät kielen ylemmillä rakennetasoilla, voivat olla tärkeitä joillakin sovellusalueilla (Rosenfeld 1995, 4), mutta niiden mallintaminen datapohjaisilla tilastollisilla menetelmillä on vaikeaa juuri niiden tilastollisen harvinaisuuden takia (Juang 1998, 40). Näiden tilastollisesti harvinaisempien kielen ylemmillä rakennetasoilla esiintyvien lingvististen ilmiöiden tai yksiköiden mallintamiseen tilastollisilla menetelmillä tarvitaan parempia ja laajempia tutkimusaineistoja kuten Sprack Jones (1994, 3) huomauttaa.

2.2. Kielimalleista

Monilla kieliteknologian sovellusalueilla lingvististä tietoa hyödynnetään nykyisin erilaisten kielimallien avulla. *Kielimalli* edustaa kielellisiä ilmiöitä abstraktiona (Edmundson 1963). Käytännössä kielimallin tehtävä on mallintaa lingvistisiä säännönmukaisuuksia, jotta niitä voidaan hyödyntää kielellisessä muodossa olevan datan (teksti tai puhe) automaattisessa prosessoinnissa (Tatham 1998, 4-17).

Kielimallit voivat olla joko *tilastollisia* tai *sääntöihin perustuvia*, joista edelliset ovat nykyisin yleisempiä kuin jälkimmäiset. Edellisistä erillisenä kielimallityyppinä voidaan vielä mainita *adaptiiviset kielimallit*, jotka edustavat suhteellisen uusia kielimallityyppejä (Clarkson 1999). Adaptiiviset (dynaamiset) kielimallit mallintavat sanojen *lyhyen aikavälin* vaihtelua erityyppisissä teksteissä tai saman tekstin eri osissa (Ueberla 1994, 32). Ne voivat myös adaptoitua sanastoon tai topiikkiin. Lingvistisesti adaptiiviset kielimallit ovat paremmin motivoituja kuin staattiset tilastolliset kielimallit, koska ne mallintavat paremmin kielen heterogeenistä luonnetta (Rosenfeld 2000a, 5). Parhaimmatkin kielimallit edustavat kielellisiä ilmiöitä vielä toistaiseksi sangen puutteellisesti (Oakes 1998, 54-55).

Yleisin tilastollinen kielimalli, jota käytetään monissa kieliteknologiasovelluksissa, on niin sanottu *n-gram* kielimalli (Rosenfeld 2000a, 3). N-gram viittaa sekvenssiin peräkkäisiä yksiköitä kuten esim. kirjaimia tai sanoja, joita voi olla n kappaletta (Oakes 1998, 253). Nykyisin käytössä olevissa n-gram kielimalleissa n on yleisesti joko kaksi (bigram) tai kolme (trigram).

Monista puutteistaan huolimatta n-gram kielimallit mallintavat tehokkaasti sanojen *lokaalista* yhdessä esiintymistä (Brill & Mooney 1997, 19) varsinkin kielissä, joissa on suhteellisen kiinteä sanajärjestys kuten esim. englannissa (Ueberla 1994, 26). N-gram kielimalleja on hyödynnetty monilla sovellusalueilla kieliteknologiassa kuten esim. automaattisessa puheentunnistuksessa, sanan sanaluokan määrittämisessä, oikeinkirjoituksen tarkistuksessa, sanan merkityksen disambiguoinnissa, tiedonhaussa ja tekstin kirjoittajan tunnistuksessa (Church 1995, 5).

N-gram kielimallit saattavat optimiolosuhteissa toimia kohtalaisen hyvin. Niiden hyvänä puolena on se, että ne voidaan opettaa helposti (Brill &

Mooney 1997, 19; Stolcke 1997, 28); toisaalta ne ovat myöskin varsin herkkiä käytetylle datalle sikäli, että pienikin muutos aineistossa aiheuttaa sen, että mallit on opetettava uudelleen - jopa samaa aihepiiriä edustavalle tekstille.

Lingvistisesti yksinkertaiset n-gram kielimallit eivät ole hyvin motivoituja, vaikka ne mallintavatkin tehokkaasti sanojen lokaalista yhdessä esiintymistä varsinkin kielissä, joissa on kiinteä sanajärjestys, kuten yllä jo todettiin. Tämä johtuu siitä (väärästä) olettamuksesta, että ainoastaan n- 1 sanaa vaikuttaisivat niitä seuraavaan sanaan (Charniak 1993, 39); todellisuudessa sanojen välisiä riippuvuuksia esiintyy varsin paljon useilla kielen eri rakennetasoilla (Aylett 2000, 25).

N-gram kielimallit eivät myöskään tyypillisesti mallinna sanojen välisiä riippuvuuksia silloin kun *n* on suurempi kuin kolme (trigrams) kuten esim. subjektin ja refleksiivipronomin välistä riippuvuutta englannin kielessä kuten esim. lauseessa "*The man* on the third floor hanged *himself* this morning".

Kaikille sana n-grammeille ei myöskään löydy aina esiintymiä laajassakaan opetusaineistossa, varsinkin jos n on suuri; tällöin joudutaan käyttämään erilaisia tekniikoita näiden esiintymättömien sana n-grammien tilastollisen todennäköisyyden arvioimiseksi (Aylett 2000, 28).

Tilastollisen kielimallin, joka tyypillisesti pyrkii rajaamaan niiden sanojen määrää, jotka voivat seurata yksittäistä sanaa lauseessa, suorituskykyä ja hyvyyttä arvioidaan informaatioteoriasta peräisin olevalla *perplexity:n* käsitteellä, joka tarkoittaa niiden vaihtoehtoisten sanojen määrää, jotka voivat seurata yksittäistä sanaa tekstissä (Lippmann 1997, 3). Kielimallin perplexity:n on todettu vaihtelevan tekstityypeittäin. Esimerkiksi radiologiaraporteissa virkerakenteet vaihtelevat vähemmän kuin englannin kielessä yleensä. Taulukossa 1 kuvataan perpelexity:ä joillakin kielenkäyttöalueilla sanelusysteemeissä, joissa käytetty sanasto vaihtelee laajuudeltaan 20000-30000 sanaan. Pienin kirjallisuudessa raportoitu perpelxity standardi Brownin korpukselle, joka sisältää miljoona amerikanenglannin sanaa, on noin 247. (Ruokos 1996.)

Eli mitä pienempi kielimallin perplexity on sovellusalueilla kuten esim. automaattinen puheentunnistus, sitä parempi se on. Kielimallin perplexity on pieni jos se on 10 tai pienempi, kohtalaisen suuri, jos se on 100 tai suurempi ja suuri, jos se on noin 200 (Zue et al. 1996).

2.3. Kielellisen tiedon käytöstä tiedonhaussa

Nykyisin lingvististä tietoa hyödynnetään kieliteknologian eri sovellusalueilla vielä toistaiseksi varsin rajallisesti. Tähän voi olla yhtenä syynä se, että (tilastollisesti harvinaisten) lingvististen ilmiöiden prosessointi erityisesti kielen ylemmillä rakennetasoilla, vaatii tietokoneelta yhä vieläkin paljon prosessointitehoa.

Sovellusalueilla kuten esim. *tiedonhaussa* tilanne lingvistisen tiedon käytön suhteen on aika tyypillinen kieliteknologiasovelluksille yleensäkin: nykyisissä (tilastopohjaisissa) hakusysteemeissä ei juurikaan hyödynnetä mitään varsinaista lingvististä prosessointia lukuun ottamatta alemman tason lingvististä prosessointia geneerisissä tehtävissä kuten sanan sanaluokan tunnistuksessa varsinkaan kielen ylemmillä tasoilla, vaan hakukyselyjen ja dokumenttien vastaavuutta määritetään yksinkertaisesti pelkästään hakusanojen avulla. Nämäkin systeemit saattavat kuitenkin itse asiassa hyödyntää implisiittisesti kielellistä tietoa, jota lingvistiseen prosessointiin pohjautuvat hakusysteemit hyödyntävät eksplisiittisesti. (Voorhees 1999, 33, 42.) Tosin kuten edellä osoitettiin tilastolliset

Taulukko 1. Joidenkin kielenkäyttöalueiden perplexity laajaa sanastoa käyttävissä sanelusysteemeissä

Domeeni	Perplexity
Radiologia	20
Journalismi	105
Yleisenglanti	247

tekniikat/kielimallit ovat usein vielä sangen puutteellisia monessakin suhteessa mitä tulee niissä hyödynnettyyn implisiittiseen tai eksplisiittiseen kielelliseen tietoon.

Esimerkkinä ylemmän tason kielellisestä tiedosta, joka voisi olla hyödyksi myös tiedonhaussa, on *pragmaattisen tiedon* hyödyntäminen puheentunnistuksessa. Esimerkiksi tunnistettavalle fraasille *speech recognition* nykyaikainen puheentunnistin voisi tunnistaa seuraavat sanat ja fraasit englannin kielessä (Steedman 1993, 243):

wreck # a # nice # beach
recognise # speech
wreck # on # ice # beach
wreck # an " eyes # peach
recondite's # beach
recon # nice # speech

Vaikka yllä oleva esimerkki on jossain määrin keinotekoinen, on se sinällään hyvinkin edustava esimerkki kielen monitulkintaisuudesta akustis-foneettisella tasolla. Tällaista monitulkintaisuutta voidaan kuitenkin vähentää huomattavasti hyödyntämällä kielellistä tietoa kielen ylemmiltä rakennetasoilta kuten esim. pragmaattista tietoa: kuten Steedman (1993, 243-244) huomauttaa *syntaktinen tieto* ei vielä riitä disambiguoimaan yllä olevia ilmauksia, koska kaikki ovat syntaktisesti koherentteja. Semanttisella tasolla monet niistä voidaan jo hylätä, mutta vasta *pragmaattisella tasolla* eli esim. kontekstissa, jossa käsitellään puheentunnistusta, kaikki paitsi tunnistettava fraasi *speech recognition* voidaan hylätä kontekstiin sopimattomina. Tiedonhaun osalta emme välttämättä tiedä vielä pragmaattisen tiedon hyödyllisyyttä tai edes sitä onko sitä edes tutkittu riittävästi juuri tällä sovellusalueella.

Alemman tason lingvistinen prosessointi, tyypillisesti sanatasolle saakka, sen sijaan näyttäisi olevan selvästi yleisempää tiedonhaussa erityisesti *geneerisissä tehtävissä* kuten esim. sanan sanaluokan määrittäminen tai sanan kantamuodon tunnistus (stemming) (Liddy 1998, 1).Esimerkkejä kielellisen tiedon hyödyntämisestä tiedonhaussa lähinnä kielen alemmilla tasoilla ovat myös esim. dokumenttien indeksointi lingvistisesti motivoitujen indeksitermien avulla (Sprack Jones 1999), hakukyselyjen laajentamiseen synonyymisten sanojen avulla (Robin & de Souza Ramalho 2001) tai paikannimien avulla (Feldman 1999, 11) sekä hakukyselyjen tulkinta tunnistamalla hakusanojen kantamuodot (Voorhees 1999, 38, 42). Kielellisen tiedon hyödyntäminen näissäkään geneerisissä tehtävissä ei useinkaan ole välttämättä helppoa: esim. synonyymilaajennuksessa vaikeutena on usein määrittää, mille sanan merkitykselle synonyymilaajennus itse asiassa tehdään, koska esim. englannin kielessä sanoilla on keskimäärin seitsemän merkitystä (Feldman 1999, 11).

Sanan merkitysten määrä ja sen suhteellinen esiintymisfrekvenssi ovat itse asiassa suhteessa toisiinsa siten, että sanan merkitysten määrä vastaa karkeasti sen suhteellisen frekvenssin neliötä (Zipf 1945) (Krovetz & Croft 1992, 124). Vaikka jonkinlainen sanan merkityksen disambiguointi tehdäänkin osana tiedonhakua, vastoin odotuksia se saattaa parantaa tuloksia vain vähän; useimmiten se saattaa jopa huonontaa tuloksia (Greengrass 2000, 87).

Yleisesti ottaen tämä johtuu siitä, että kyselyiden monimerkityksisten hakusanojen disambiguointi voidaan tehdä kyselyn muiden hakusanojen avulla: esim. jos kyselynä on *bank* AND *economy*, niin hakutuloksista ei yleensä löydy dokumentteja, joissa sana bank esiintyy merkitykessä 'rantapenger' (Krovetz & Croft 1992). Tilanne on tietysti toinen, jos hakusanana on pelkästään sana *bank*.

Voorhees:n (1999, 41-45) mukaan seuraavia johtopäätöksiä voidaan tehdä kielellisen tiedon hyödyntämisestä tiedonhaussa tällä hetkellä:

(1) Kielellistä tietoa hyödyntävien tekniikoiden täytyy olla *lähes täydellisiä*, ennen kuin niistä on apua käytännössä. Koska alemman tason lingvististä prosessointia suorittavat tekniikat kuten esim. sanan sanaluokan määrittäminen tai sanan merkityksen disambiguointi eivät toimi vielä täydellisesti, virhemarginaalin vaikutus on otettava huomioon tuloksissa.

(2) *Hyvien hakukyselyjen merkitys* on tärkeä tiedonhaussa yleensäkin ja erityisen tärkeä se on jonkinlaiseen lingvistiseen prosessointiin perustuvissa tekniikoissa; liian lyhyet hakukyselyt eivät tarjoa paljonkaan edellytyksiä minkäänlaiselle lingvistiselle prosessoinnille, esim. luonnollisen kielen prosessointiin perustuvissa tekniikoissa, jotka toimivat paremmin juuri pitemmillä hakukyselyillä kuten osoitettiin TREC-5:ssä. Hakukyselyiden pitäisi myös sisältää indeksitermejä, joita dokumenteissa käytetään. Valitettavasti vain avainsanat (indeksitermit) edustavat dokumentissa käsiteltyjä topiikkeja jossain määrin epäluotettavasti; on mahdollista

puhua tietystä topiikista käyttämättä yhtä ainoa topiikkisanaa.

(3) Koska tilastolliset tekniikat jo hyödyntävät *implisiittisesti* kielellistä tietoa, lingvistiseen tietoon perustuvat tekniikat eivät välttämättä paranna hakutuloksia paljoakaan, vaikka ne toimisivatkin täydellisesti, koska molemmat tekniikat hyödyntävät jo itse asiassa samaa tietoa, toiset implisiittisesti toiset taas eksplisiittisesti.

(4) Termien normalisaatio, jossa saman sanan eri kirjoitusasuille, esim. *on-line, on line*, määritetään yksi kantamuoto voi olla hyödyllisempää kuin sanan kantamuodon tunnistus ns. stemming-tekniikan avulla erisnimien käsittelyn lisäksi.

Alemman tason lingvististä prosessointia hyödyntävien tekniikoiden puutteellisuuksien lisäksi *virheellisten hakukyselyjen*, joissa voi esiintyä kirjoitusvirheitä, neologismeja tai muita kielivirheitä, vaikutus tiedonhaun tuloksiin täytyy myös aina ottaa huomioon.

Hakukyselyjen virheellisyyksien lisäksi virheitä esiintyy myös prosessoitavassa datassa itsessään sekä tekstissä että puheessa. Tietokantakyselyjen virheitä koskevissa tilastollisissa tutkimuksissa, esim. Eastman & McLean (1981), Thompson (1980) huomattiin, että jopa 25-30%:a kirjoitetuista hakusanoista oli virheellisiä (Bates et al. 1993, 25), mikä tuntuisi korostavan virheenkorjaustekniikoiden tarvetta kieliteknologiasovelluk sissa. Sinällään virheen tunnistus on helpompaa kuin virheen korjaaminen.

3. Kielellisen tiedon laajamittaisempi hyödyntäminen tiedonhaussa

Vaikka tiedonhaku on eräs keskeisimpiä kieliteknologian sovellusalueita, jossa merkityksen identtisyys hakukyselyjen ja etsittävien dokumenttien välillä on sen keskeisin käsite (Jurafsky & Martin:n 2000, 600), käytännön sovellukset operoivat lähes yksinomaan *sanatasolla*. Tämä johtunee osaltaan siitä, että kieliteknologian eri sovellusalueilla potentiaalisesti tai tosiasiallisesti hyödyllisen kielellisen tiedon on aina lähtökohtaisesti sovelluttava tietokoneympäristöön.

Keskeinen ongelma tiedonhaussa tällä hetkellä on hakujärjestelmien riittämätön *suorituskyky*: paljon aikaa kuluu hukkaan etsittäessä relevantteja dokumentteja suuresta määrästä epärelevantteja dokumentteja samalla kun osa relevanteista dokumenteista jää löytymättä.

Kielellistä tietoa, joka on ollut tähän saakka menestyksellisintä sovellusalueilla kuten esim. tiedonhaku, voidaan luonnehtia yleensä ottaen jollain tapaa *rajalliseksi* (finate) kuten esim. sanan sanaluokkien määrä, temaattiset roolit tai semanttiset primitiivit (Jurafsky & Martin 2000, 616). Tämä ei sinänsä vielä tarkoita sitä, etteikö tällainenkin tieto, kuten esim. yksittäisen sanan sanaluokka, olisi monitulkintainen kielissä kuten esim. englannin kieli. Tällöin siis yksittäisellä sanalla kuten esim. sanalla *can* voi olla useita eri sanaluokkia. Tällaisen jollakin tapaa rajallisen kielellisen tiedon merkitystä sovellusalueilla kuten esim. tiedonhaussa kannattaisi ehkä tutkia lisääkin. Perinteisen kielellisen tiedon, jota siis voidaan kuvata osittain kielen rakennetasomallin avulla, ohella joidenkin *kielellisten intuitioiden*, kuten esim. sen että samaa sanaa käytetään tyypillisesti samassa merkityksessä yksittäisen dokumentin sisällä, hyödyntäminen on osoittautunut hyödylliseksi sovellusalueilla kuten esim. tiedonhaussa. Näitä kielellisiä intuitioita ja niiden hyödyllisyyttä tiedonhaussa kannattaisi ehkä tutkia lisää. Usein näitä kielellisiä intuitioita kuvataan tutkimuksissa ilmauksella "On yleisesti tunnettua, että …".

On vaikea antaa esimerkkejä sellaisesta kielellisestä tiedosta tiedonhaussa, jota ei olisi absoluuttisesti testattu tai käytetty missään aiemmin. Kielellinen tieto saattaa parantaa hakutuloksia joissakin tapauksissa mutta ei välttämättä kaikissa tapauksissa, esim. kieliopillisesti rakennetut tesaurukset ovat parantaneet tiedonhaun tuloksia lyhyille dokumenteille, kuten esim. kuvien kuvatekstit, mutta pidemmissä tekstidokumenteissa niistä ei ole ollut mainittavaa hyötyä (Walker 2001, 9-10).

Eräs ajatus, jota kannattaisi kokeilla topiikkisanastojen muodostamisessa, on monikielisten sanastojen hyödyntäminen: esim. topiikkiin auto liittyy englannin kielessä fraaseja, jotka sisältävät sanan *car* kuten esim. *car park, car accident*, etc., jotka löydetään sellaisenaan fraasisanakirjoista. Löytyy kuitenkin myös suuri määrä topiikkisanoja, joissa ei esiinny sanaa *car* kuten esim. *garage* tai *motor trip,* jotka myös liittyvät samaan autotopiikkiin.

Näiden topiikkisanojen tunnistamiseen voitaisiin käyttää suomi-englanti sanakirjaa, siten että hakusanalle auto etsittäisiin käännösten avulla sen topiikkisanoja englannin kielessä kuten esim. *garage* 'autotalli' tai *motor trip* 'automatka'. Erikielisillä sanastoilla saattaisi olla erilainen

kuvausvoima topiikkisanastojen luomisessa siten, että voitaisiin ehkä luoda topiikkisanastoja jopa monen eri kielen kautta. Sanoille, joilla on useita mahdollisia käännöksiä, voitaisiin käyttää yleisesti hyväksi havaittua tekniikkaa valita tilastollisesti yleisin merkitys.

Ylläkuvattu idea vaikuttaa sinänsä niin yksinkertaiselta, että sitä on erittäin todennäköisesti jo kokeiltu tiedonhaussa. Se missä sitä ei ole ehkä vielä huomattu kokeilla on tekstinennakointiohjelmistot (word prediction software), joiden tarkoituksena on nopeuttaa tekstin tuottamista (typing) ennustamalla sana, jota kirjoittaja kirjoittaa joko parhaillaan (ns. word completion) tai aikoo kirjoittaa seuraavaksi (ns. word prediction). Esim. sanan *speech* jälkeen kirjoittaja voisi kirjoittaa seuraavaksi sanaksi sanan *recognition* puheentunnistusta käsittelevässä tekstissä. Tällöin sovellus voi siirtää suoraan sanan *recognition* osaksi tekstiä jollain näppäinkomennolla, esim. välilyöntikomennolla. Kun käyttäjä on kirjoittanut esim. sanan car, niin tekstinennakointisovellus voisi yksinkertaisimmillaan ennustaa seuraavaksi kirjoitettaviksi samaa aihetta käsitteleviksi topiikkisanoiksi sanoja tai fraaseja, jotka sisältävät sanan *car*, esim. *car park*, *car accident*, etc. Kuten edellä todettiin, on myös muita fraaseja, jotka eivät sisällä sanaa *car*, joita voitaisiin kuitenkin ennustaa ainakin jossain märin kaksikielisen sanakirjan avulla kuten yllä kuvattiin. Vielä toistaiseksi tekstinennakointiohjelmistot toimivat yksikielisesti lähinnä lausetasolla. Yllä kuvattu esimerkki kuvaa erästä lingvistisen tiedon hyödyntämistapaa eli jo olemassa olevan tekniikan hyödyntämistä täysin uudella sovellusalueella.

Mitä sitten kielellisen tiedon laajamittaisempi hyödyntäminen kieliteknologiasovelluksissa vaatisi käytännössä? Ainakin seuraavia toimenpiteitä voisi harkita (vrt. Väyrynen 2002):

(1) Analyysi potentiaalisesti hyödyllisestä lingvistisestä tiedosta siten kuin sitä on hyödynnetty joillakin keskeisillä kieliteknologian sovellusalueilla tällä hetkellä.

Tässä analyysissä voitaisiin ottaa huomioon lingvistisen tiedon *kielestä riippumattomuus/ kielikohtaisuus*, lingvistisen tiedon *lingvistinen motivoitavuus* (mikä sinällään ei välttämättä aina paranna sovellusten suorituskykyä varsinkaan tilastollisissa menetelmissä Rosenfeld 2000b, 1321).

(2) Karkeiden valintakriteerien määrittämien potentiaalisesti hyödylliselle lingvistiselle tiedolle suhteessa erilaisiin ohjelmisto-ominaisuuksiin kuten esim. ohjelman robusti toiminta tai sovellukselta vaadittava suoritustaso.

(3) Perusohjelmistokomponenttien (modules, semiproducts, ks. Cucchiarini et al. 2001; Väyrynen 2002) kartoitus osana digitaalista infrastruktuuria ja hyödyntäminen eri sovellusalueilla.

(4) Yhteistyö sisällön tuottajien ja sovellusten kehittäjien välillä sovellusten suorituskyvyn parantamiseksi.

(5) Automaattinen lingvistinen indeksointi kaikilla kielen rakennetasoilla. Vaikka tätä indeksointia, joka voitaisiin tuottaa esim. tekstinennakointiohjelmiston (word prediction software) avulla automaattisesti, ei voitaisikaan vielä hyödyntää täysimääräisesti, se olisi kuitenkin valmiina myöhempää käyttöä varten.

(6) Kielellisen metadatan laajamittaisempi hyödyntäminen esim. tekstinennakointisovelluksen tuottama tekstin kielellisten piirteiden metadatakuvauksena.

(7) Rajoitettujen kielimuotojen (controlled languages, sublanguages) käyttäminen tekstintuottamisessa hyödyntäen domeenin käsitettä (ks. esim. Sekine 1998).

(8) Eri datamuotojen yhteisanalyysi, jolla voidaan korvata yksittäisen datamuodon analyysin puuteita, silloin niitä on saatavilla kuten esim. DVD-formaatissa audion ja tekstianalyysin hyödyntäminen sisällön kuvailussa ja tiedonhaussa.

4. Päätäntä

Lingvistisen tiedon laajamittaisempaan hyödyntämiseen kieliteknologian eri sovellusalueilla ei sinällään liene mitään patenttiratkaisua, vaan lingvistisen tiedon hyödyllisyys saattaa määräytyä pitkälle sovellusalueen ja käytetyn kielen mukaan.

Monilla sovellusalueilla kuten esim. puheentunnistuksessa sovelluksen suorituskyky on itse asiassa monien tekijöiden summa, lingvistinen tieto vain yhtenä niistä. Tiedonhaun osalta katsotaan usein, että noin 70%:n tarkkuus olisi jo todella hyvä suoritustaso.

Osalla kohdan kolme alla luetelluista toimenpiteistä, esim. rajoitettujen kielimuotojen käytöllä, on sekä etuja että haittoja käytännössä: sanastoltaan ja lauserakenteeltaan rajoitettu kielimuoto parantaa sinällään tekstin luettavuutta (mikä voisi olla yhtenä lingvistisenä metadatapiirteenä

analysoituna esim. perinteellisillä tekstin luettavuutta määrittävillä ohjelmistoilla) ja harmonisoi tuotetun tekstin laatua kauttaaltaan, mutta monista rajoituksista johtuen kirjoittajalle tällaisen tekstin kirjoittaminen on varsin hankalaa (Hoard 1998, 227). Siitä huolimatta jotkut tutkijat, kuten esim. Hoard (1998, 230), ovat kuitenkin sitä mieltä, että täysin *automaattinen konekäännös* on mahdollista vain hyödyntämällä jotain rajoitettua kielimuotoa, jolla käännettävä teksti tuotetaan.

Monilla sovellusalueilla kuten esim. automaattisessa puheentunnistuksessa perinteellinen kielen rakennetasoihin perustuva lingvistisen tiedon kuvausmalli on korvattu uudella kielen rakennemallilla, jossa esim. kielen fonologista tasoa kuvataan ääntämyssanakirjan avulla ja syntaktista tasoa sana n-grammeilla (Huckvale 1996, 9).

Monilla kieliteknologian sovellusalueilla luonollisen kielen prosessointiin katsotaan liittyvän eniten potentiaalia mitä tulee tehokkuuden parantamiseen. Vielä toistaiseksi luonnollisen kielen prosessointi ei ole täysin lunastanut suurta lupaustaan sovellusalueilla kuten esim. tiedonhaussa (Harman et al. 1996), vastakkaisista esimerkeistä huolimatta (ks. Woods et al. 1999).

Kielellisten resurssien merkitys on suuri kieliteknologiassa ja näyttääkin siltä, että kieliresursseilla, kuten esim. tesaurustyyppisillä sanakirjoilla ja erisnimiä sisältävillä leksikoilla, on vielä toistaiseksi *suurempi merkitys* tiedonhaussa kuin luonnollisen kielen prosessoinnin tekniikoilla.

Tähän on osasyynä se, että luonnollisen kielen prosessoinnin tekniikat eivät pääsääntöisesti pysty prosessoimaan suuria tekstimassoja domeenista riippumatta. (Harman et al. 1996.)

Systeemien, jotka pohjautuvat puhtaasi lingvistiseen prosessointiin useilla kielen rakennetasoilla kuten esim. DR-LINK (joka on luonnollisen kielen prosessointia kaikilla kielen tasoilla soveltava tiedonhakusysteemi, lisätietoa: www.textwise.com tai www.mnis.net Liddy 1998, 4), kehittäminen on usein kallista ja aikaa vievää varsinkin jos sovellukset kehitetään täysin "puhtaalta pöydältä" hyödyntämättä olemassa olevaa teknologiaa. Tämä ei kuitenkaan saisi estää meitä kun yritämme kuroa umpeen semanttista kuilua tiedontarvitsijan ja bittivirran välillä.

Hyväksytty julkaistavaksi 18.6.2002.

Viitteet

Aylett, M. P. (2000). Stocastic Suprasegmentals: Relationships between Redundancy, Prosodic Structure and Care of Articulation in Spontaneous Speech. Ph.D. thesis. University of Edinburgh.

Bates, M. Borrow, R. J., Weischedel, R. M (1993). Critical Challenges for Natural Language Processing. Challenges in Natural Language Processing. Toimittaneet Bates, M. & Weischedel, R. M. Cambridge University Press.

Brill, E. & Mooney, R. J. (1997). An Overview of Empirical Natural Language Processing. AI Magazine. Vol. 18. No.4:13-24.

Charniak, E. (1993). Statistical Language Learning. Cambridge Mass: The MIT Press.

Clarkson, P. R. (1999). Adaptation of Statistical Language Models for Automatic Speech Recognition. Ph.D. thesis. Darwin College. University of Cambridge and Cambridge University Engineering Department.

Crystal, D. (1997). The Cambridge Encyclopaedia of Language. Cambridge: Cambride University Press.

Cucchiarini, C., Daeleman, W., Strik, H. (2001). Srengthening the Duch Human Language Technology Interface. The ELRA Newsletter: 3-7.

Eastman, C. M. & McLean, D. S. (1981). On the Need for Parsing Ill-formed Input. American Journal of Computational Linguistics 7(4):257.

Edmundson, H. P. (1963). A Statistician's Linguistic Models and Language Data Processing. Natural Language and the Computer. Toimittanut Gavin, P. L. New York: McGraw Hill.

Feldman, S. (1999). NLP Meets the Jabberwocky: Natural Language Processing in Information Retrieval. 5th of April 2001. http://www.onlineinc.com/onlinemag/OL1999/feldman5.html.

Greengrass, E. (2000). Information Retrieval: A Survey. 22nd of Februaray 2002. http://www.cs.umbc.edu/cadip/readings/IR.report.120600.book.pdf.

Harman, D., Schäuble, P., Smeaton, A. (1996). The Survey of the State of the Art in Human Language Technology. Toimittanut Cole, R. 27th of March 2000. http://cslu.cse.ogi.edu/HLTsurvey.

Hoard, J. E. (1998). Language Understanding and the Emerging Alignment of Linguistics and Natural Language Processing. Using Computers in Linguistics. A Practical Guide. Toimittaneet Lawler, J. & Dry, H. A. London and New York: Routledge.

Huckvale, M. (1996). Learning from the Experience of Building Automatic Speech Recognition Systems. UCL Working Papers, Speech, Hearing and Language.

Juang, B. H. (1998). The Past, Present, and Future of Speech Processing. In IEEE Signal Processing Magazine. Vol. 15. No. 3:24-48.

Jurafsky, D. & Martin, J. H. (2000). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice-Hall.

Krovetz, R. & Croft, W. B. (1992). Lexial Ambiguity in Information Retrieval". ACM Transactions on Information Systems. Vol. 10. No.2:115-141.

Liddy, E. D. (1998). Enhanced Text Retrieval Using Natural Language Processing. 9th of February 2001. http://www.asis.org/Bulletin/Apr~98/liddy.html.

Lippmann, R. P. (1997). Speech Recognition by Machines and Humans. Speech Communication 22: 1?15.

Oakes, M. P. (1998). Statistics for Corpus Linguistics. Edinburgh University Press.

Robin, J. & de Souza Ramalho, F. (2001) Empirically Evaluating WordNet-Based Query Expansion in a Web Search Engine Setting. IR2001. Toimittanut Timo Ojala.

Rosenfeld, R. (1995). Optimising Lexical and N-gram Coverage Via Judicious Use of Linguistic Data. In Proc. Eurospeech'95.

Rosenfeld, R. (2000a). Two Decades of Statistical Language Modeling: Where Do We Go from Here?" In Proceedings of the IEEE 88(8).

Rosenfeld, R. (2000b). Incorporating Linguistic Structure in Statistical Language Models. Philosophical Transactions of the Royal Society. Series A, 358 (1769): 1311-1324.

Ruokos, S. (1996). Language Reprentation. The Survey of the State of the Art in Human Language Technology. Toimittanut Cole, R. 27th of March 2000. http://cslu.cse.ogi.edu/HLTsurvey.

Schmandt, C. (1994). Voice Communication with Computers. New York. Van Nostrand Reinhold.

Sekine, S. (1998). Corpus Based Parsing and Sublanguage Studies. Ph.D. thesis. Computer Science Department. New York University.

Sprack Jones, K. (1994). Natural Language Processing: She Needs Something Old and Something New (Maybe Something Borrowed and Something Blue, Too). Presidential Address, June 1994. Association for Computational Linguistics.

Sprack Jones, K. (1999). What is the Role of NLP in Text Retrieval? Natural Language Information Retrieval. Toimittanut Strzalkowski, T. Dordrecht/Boston/London:Kluver Academic Publishers.

Steedman, M. (1993). "Surface structure, intonation, and discourse meaning". In Bates, M. & Weischedel, R. M. (eds.) 1993.

Stolcke, A. (1997). Linguistic Knowledge and Empirical Methods in Speech Recognition. AI Magazine. Vol. 18. No. 4: 25-31.

Tatham, M. (1998). Teaching Notes. 1st of October 2001. www.essex.ac.uk/speech/teaching/lg433-98-17.html.

Thompson, B. H. (1980). Linguistic Analysis of Natural Language Communication with Computers. Proceedings of the Eighth International Conference on Computational Linguistics:190-201.

Ueberla, J. (1994). Analyzing and Improving Statistical Language Models for Speech Recognition. Ph.D. thesis. Simon Fraser University.

Ullman, S. 1966. Language and Style. Oxford: Basic Blackwell.

Voorhees, E. M. (1999). Natural Language Processing and Information Retrieval. Information Extraction: Towards Scalable, Adaptable Systems: 32-48. Toimittanut Pazienza, M. T. Germany: Springer.

Väyrynen, P. (2002). "Kohti digitaalista infrastruktuuria kieliteknologiassa". In Informaatiotutkimus 1/2002.

Walker, D. (2001). "Query expansion using thesauri: previous approaches and possible new directions". In IS-242: Information Retrieval Systems.

Woods, W. A., Bookman, L. A., Houston, A., Kuhuns, R. J., Martin, P., Green, S. (1999). Linguistic Knowledge Can Improve Information Retrieval. SMLI TR-99-83. 21st of February 2002. http://research.sun.com/techrep/1999/smli_tr-99-83.pdf.

Zipf, G. (1945). The Meaning-Frequency Relationship of Words". J Gen Psycho 33:251-266.

Zue, V., Cole, R., Ward, W. (1996). Speech Recognition. The Survey of the State of the Art in Human Language Technology. Toimittanut Cole, R. 27th of March 2000. http://cslu.cse.ogi.edu/HLTsurvey.