Luonnollisen kielen tilastollinen käsittely

T-61.281 (3 ov) L

Kevät 2003

Luennot: **Timo Honkela**
Laskuharjoitukset: **Vesa Siivola**

Luentokalvot: Krista Lagus ja Timo Honkela

1. Tilastollinen konekääntäminen

- Automaattinen kielenkääntäminen on eräs pitkäaikaisimmista kielioteknologian tavoitteista.
- Konekääntäminen (machine translation) on kuitenkin hyvin vaikea ongelma.
- Nykyisten konekäännösohjelmien tulos toimii lähinnä raakakäännöksenä, joka voi nopeuttaa aidon kielenkääntäjän työtä, muta ei sellaisenaan kelpaa ihmislukijalle.
- Hyvin rajallisissa sovellusalueissa (kuten säätiedotukset) voidaan päästä kohtuulliseen lopputulokseen täysin automaattisesti; Kanadassa englanti-ranska -käännös ja Suomessa suomi-ruotsi -käännös.
- HAMT = Human Aided Machine Translation, MAHT = Machine Aided Human Translation, L10N = localisation

Kääntämisen eri tasoja

- Yksinkertaisin lähestymistapa on *sanasta sanaan käännös* korvaa lähtökielen sanoja kohdekielen sanoilla. Lopputuloksen sanajärjestys on usein väärä.

- *Muunnosmenetelmät* (syntaktinen ja semanttinen) rakentavat rakenteisen välirepresentaation lähtökielen sanajonosta muuntavat sen kohdekielen välirepresentaatioksi (jonkinlaisia sääntöjä käyttäen) ja generoivat tästä kohdekielen sanajonon.

- *Syntaktinen muunnosmenetelmä* rakentaa lähtökielen sanajonosta syntaktisen rakennekuvauksen. Lähestymistapa edellyttää toimivaa syntaktista disambiguointia.

  Tällä tavoin voidaan ratkaista sanajärjestysongelmat, mutta usein lopputulos ei ole semanttisesti oikein. Esim. saksan 'Ich esse gern' (Syön mielelläni) kääntyisi syntaktisella menetelmällä 'I eat readily' (tai 'willingly', 'gladly'). Englannissa saksan ilmausta vastaavaa verbi-adverbiparia ei kuitenkaan ole, vaan oikea käännös olisi 'I like to eat'.

- *Semanttisissa muunnosmenetelmissä* tehdään syntaktista jäsennystä täydellisempi kuvaus, semanttinen jäsennys, jonka tarkoituksena on saada aikaan käännös, joka on myös semanttisesti oikein.

  Kuitenkin semanttisesti 'sanatarkka' käännös voi olla kohdekielessä kömpelö, vaikka onkin periaatteessa ymmärrettävissä. Esim. espanjan lauseen 'La botella entró la cueva flotando' tarkka käännös olisi 'the bottle entered the cave floating' (pullo tuli luolaan kelluen) mutta luontevampaa olisi sanoa 'the bottle floated into the cave' (pullo kellui luolaan).

  Useiden kömpelöiden ja epäluontevien käännösten käyttö hidastaa ymmärtämistä, vaikka ymmärtäminen olisikin periaatteessa mahdollista. Monitulkintaisuuden mahdollisuudesta johtuen epäluonteva käännös voidaan myös helpommin tulkita väärin.

- *Interlingua* – keinotekoinen yleinen (kieliriippumaton) välikieli tai tietämysrepresentaatio. Käännetään lähtökielestä interlingualle ja interlinguasta mille tahansa kohdekielelle.

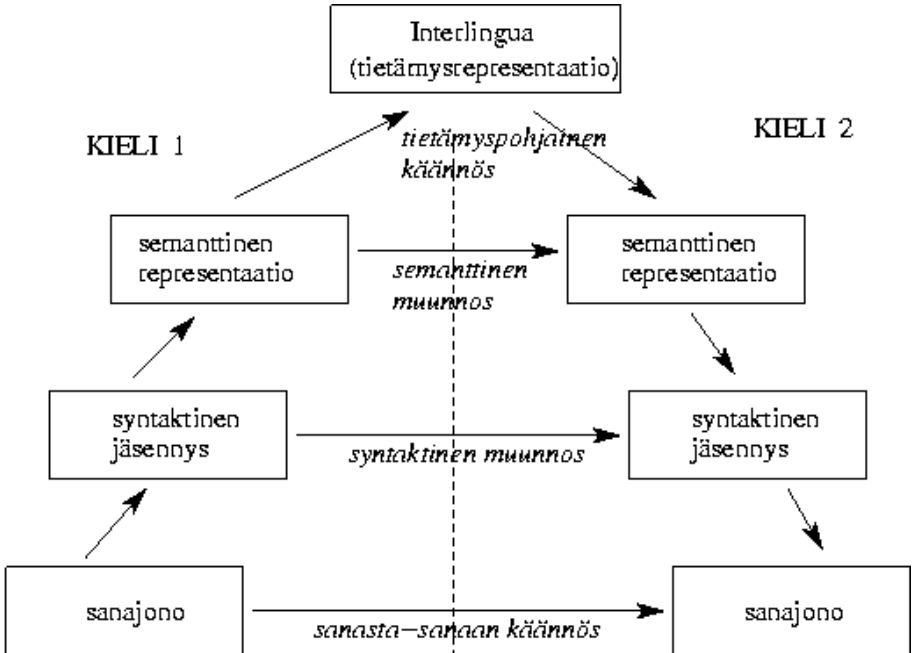
  Kääntimiä n kielen välille tarvitaan tällöin n^2 kpl sijaan vain $2n$ kpl. Lisäksi ne voidaan toteuttaa mahdollisimman suurelta osin yleiskäyttöisillä kielenkäsittelymenetelmillä. Kuitenkin riittävän välikieler määrittely on itsessään hankala ongelma, jota ei ainakaan toistaiseksi ole ratkaistu riittävässä laajuudessa.

Seuraavan sivun kuvassa on näytetty konekäännösjärjestelmän vaihtoehtoiset toteutustavat.

Tilastollisen kielenkäsittelyn menetelmiä voidaan käyttää järjestelmän komponentteina minkä tahansa nuolen kohdalla (esim. jäsentäminen, disambiguointi jne).

Konekääntimet voivat myös olla kombinaatioita symbolisista ja tilastollisista komponenteista.

Pelkästään kielenkääntämiselle erityinen ongelma on *tekstinlinjaus*.


Interlingua
(tietämysrepresentaatio)
KIELI 1
KIELI 2
tietämyspohjainen
käännös
semanttinen
representaatio
semanttinen
muunnos
semanttinen
representaatio
syntaktinen
jäsennys
syntaktinen muunnos
syntaktinen
jäsennys
sanajono
sanasta–sanaan käännös
sanajono

1.1 Tekstinlinjaus

- Tekstinlinjauksella (text alignment) tarkoitetaan kahden erikielisen *rinnakkaistekstin* asettamista kohdakkain siten, että osoitetaan toisiaan vastaavat tekstijonot.

- Rinnakkaisteksteillä tarkoitetaan saman dokumentin erikielisiä käännöksiä.

- Useimmin käytetyt rinnakkaistekstit ovat hallinnollisia tekstejä peräisin maista tai valtioliitoista, joissa on useita virallisia kieliä (esim. EU, Kanada, Sveitsi, Hong Kong).

- Helpon saatavuuden lisäksi hallinnolliset rinnakkaistekstit ovat yleensä konsistentisti ja mahdollisimman tarkasti käännettyjä. Tällainen aineiston korkea laatu on tärkeää sekä tilastollisten menetelmien kehittämiselle että menetelmien evaluoinnille.

- Myös sanoma- ja aikakauslehtiä joskus käytetään, ja myös uskonnollisia tekstejä olisi helposti saatavilla. Kuitenkin tulokset ovat yleensä selvästi heikompia, oletettavasti johtuen vähemmän sanatarkoista ja konsistenteista käännöksistä, ja vähemmän stationaarisesta tekstilajista (esim. ajankohtaiset uutisaiheet muuttuvat nopeasti).

- Tekstinlinjauksessa on yleensä kaksi vaihetta:

    1. Lauseiden ja kappaleiden linjaus: tekstin raakalinjaus, jossa toisiaan vastaavat kappaleet, lauseet ja lauseparit asetetaan suunnilleen kohdakkain.
    2. Sanojen linjaus ja kaksikielisen sanakirjan indusointi, jossa raakalinjatun aineiston perusteella etsitään lähdekielisiä sanojen (ja fraasien) kohdekieliset vastineet.

1.2 Lauseiden ja kappaleiden linjaus

Yleensä lauseiden linjaus on välttämätön ensimmäinen askel monikielisen korpuksen tuottamisessa.

Konekäännöksen ja kaksikielisten sanakirjojen tuottamisen lisäksi linjaus voi hyödyttää myös muita sovelluksia kuten

- Sananmerkitysten disambiguointi: sanan eri merkityksiä voidaan ryhmitellä sen saamien eri käännösvastineiden perusteella.
- Monikielinen tiedonhaku: Tiedonlähde voi olla eri kielellä kuin millä kysymys esitetään.
- Kääntäjän apuväline: Kun dokumenttien tiedot muuttuvat, voidaan automaattisesti osoittaa toisenkielisen dokumentin kohta, jota täytyy myös päivittää, ja ehkä ehdottaa päivitystä.

Jyvitys

Jyvä (bead) on lause tai muutaman lauseen jono ja sitä vastaavat (linjatun tekstin) toisenkielinen lausejono. Kumpi tahansa jono voi olla myös tyhjä. Jokainen lause kuuluu täsmälleen yhteen jyvään.

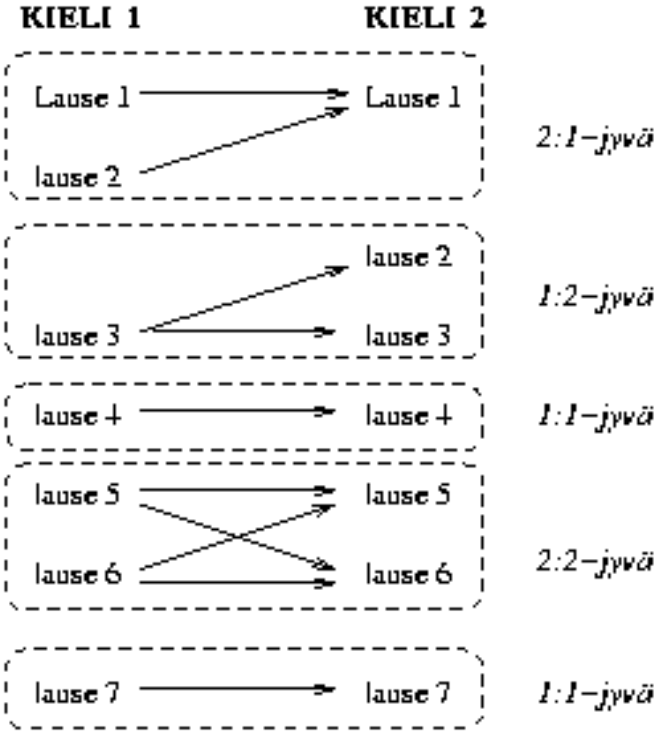
Jyvitys on kuvaus, jossa tekstit on jaettu osiin ja kerrottu, mitä kielen 1 osaa mikäkin kielen 2 osa vastaa.

Lauseiden linjaus ei ole triviaali ongelma, koska yhtä lähtökielen lausetta ei läheskään aina vastaa yksi kohdekielen lause (1:1-jyvä).

1:2 ja 2:2-jyvät (myös 1:3 ja 3:1): Lauseita pilkotaan eri tavoilla. Ihmiskääntäjä käyttää eri järjestyksiä tehdäkseen lopputuloksesta luontevan.

2:2-vastaavuudessa lähtökielen kahden peräkkäisen lauseen osia esitetään kohdekielen kahdessa peräkkäisessä lauseessa (riittävä päällekkäisyys).

Milloin päällekkäisyys on riittävä? Yleensä muutaman sanan siirtyminen ei riitä, vaan edellytetään kokonaisen lausekkeen päällekkäisyyttä.


KIELI 1
KIELI 2
Lause 1
Lause 1
lause 2
2:1–jyvä
lause 2
lause 3
lause 3
1:2–jyvä
lause 4
lause 4
1:1–jyvä
lause 5
lause 5
lause 6
lause 6
2:2–jyvä
lause 7
lause 7
1:1–jyvä

Poistot ja lisäykset eli 1:0 ja 0:1-jyvät:

Joitain asioita voidaan sanoa eksplisiittisesti toisella kielellä mutta jättää pois toisella kielellä, koska ne oletetaan implisiittisesti tulkittaviksi (ehkä asioiden erilaisen järjestyksen ansiosta, ehkä sanojen erilaisten sivumerkitysten takia, ehkä kulttuurisista syistä).

Eri tutkimusten perusteella voidaan arvioida, että n. 90% vastaavuuksista on tyyppiä 1:1 (tosin osuus on luultavasti kielipari- ja tekstilajiriippuva).

On myös melko tavallista, että kääntäjät järjestävät lauseita eri järjestyksiin. Tässä esitetyt mallit eivät kuitenkaan kykene representoimaan tätä mahdollisuutta vaan tulkitsevat tapaukset mm. poistoiksi ja lisäyksiksi.

Tekstinlinjauksen tilastollisia menetelmiä

Osa tilastollisista menetelmistä perustuu ainoastaan tekstinpätkien pituuksien tarkasteluun, osa taas huomioi lauseissa käytetyn sanaston (merkkijonot).

- Tekstinpätkien pituuksiin perustuvat menetelmät
- Identtisiin merkkijonoihin perustuva menetelmä
- Leksikaaliset menetelmät

Jatkossa: olkoon kielen 1 teksti S jono lauseita $S = (s_1, \ldots, s_I)$ ja kielen 2 teksti T samoin $T = (t_1, \ldots, t_J)$ (S = source, T = target)

Tekstinpätkien pituuksiin perustuvat menetelmät

Useat varhaiset tekstinlinjausmenetelmät ovat tätä tyyppiä.

Etsitään linjaus A, jolla on suurin tn:

$$\arg\max_A P(A|S,T) = \arg\max_A P(A,S,T) \tag{1}$$

(todennäköisin linjaus voidaan etsiä mm. dynaamisella ohjelmoinnilla).

Useat menetelmät jakavat linjatun tekstin jonoksi jyviä $(B_1, \ldots, B_K)$ ja approksimoivat koko linjatun tekstin tn:ää olettamalla, että jyvän tn ei riipu sen ympäristöstä, vaan ainoastaan jyvän sisältämistä lauseista:

$$P(A,S,T) = \prod_{k=1}^{K} P(B_k) \tag{2}$$

Jyvän todennäköisyyden laskenta

Gale & Church, 1991, 1993:

Jyvän tn. riippuu jyvässä olevien lauseiden pituuksista merkkeinä mitattuna. Perustuu oletukseen, että yhden kielen pitkiä pätkiä todennäköisesti vastaavat pitkät pätkät myös toisessa kielessä.

Oletetaan, että aineistot on jo linjattu kappaletasolla (laskennallisen tehokkuuden vuoksi).

Sallitaan vain linjaustyypit $\{1:1, 1:0, 0:1, 2:1, 1:2, 2:2\}$

Olkoon $D(i,j)$ etsitty pienimmän kustannuksen linjaus lauseiden $s_1, \ldots, s_i$ ja $t_1, \ldots, t_j$ välillä.

Lasketaan $D(i,j)$ rekursiivisesti. Perustapaus, määritellään: $D(0,0) = 0$.

Rekursio:

$$
\begin{aligned}
D(i,j) = \min \quad & D(i,j-1) && + cost(0:1 \text{ linjaus } 0, t_j) \\
& D(i-1,j) && + cost(1:0 \text{ linjaus } s_i, 0) \\
& D(i-1,j-1) && + cost(1:1 \text{ linjaus } s_i, t_j) \\
& D(i-1,j-2) && + cost(1:2 \text{ linjaus } s_i, t_{j-1}, t_j) \\
& D(i-2,j-1) && + cost(2:1 \text{ linjaus } s_{i-1}, s_i, t_j) \\
& D(i-2,j-2) && + cost(2:2 \text{ linjaus } s_{i-1}, s_i, t_{j-1}, t_j)
\end{aligned}
$$

Kunkin tyyppisen linjauksen (jyvän) kustannus lasketaan seuraavasti:

Oletetaan malli: yksi kielen L_1 merkki generoi satunnaisen määrän merkkejä kieleen L_2. Oletetaan generoinneille gaussinen tn-jakauma, jonka keskiarvo μ ja varianssi σ^2 estimoidaan suurista rinnakkaiskorpuksista (saksa/englanti-parille estimoitiin $\mu = 1.1$ koko korpuksesta, ranska/englanti-parille 1.06. Varianssi voidaan estimoida kappaletason linjausta hyväksikäyttäen.)

Kustannuksena voidaan käyttää tekstinpätkien etäisyyden negatiivista log-

likelihoodia mallissa:

$$cost(l_1, l_2) = -\log P(\alpha \text{ linjaus } |\delta(l_1, l_2, \mu, \sigma^2) \quad (3)$$

jossa α on jokin sallituista linjaustyypeistä ja $\delta(l_1, l_2, \mu, \sigma^2) = (l_2 - l_1\mu)/\sqrt{l_1\sigma^2}$

Tarvittavat todennäköisyydet estimoidaan soveltamalla Bayesin kaavaa

$$P(\alpha|\delta) = P(\alpha)P(\delta|\alpha) \quad (4)$$

Tällöin siis 1:1-linjauksen suuri prioritodennäköisyys (90 %) aiheuttaa preferenssiä sen valintaan.

Rekursiivinen kustannusten laskenta-algoritmi on hidas, jos tekstinpätkät ovat pitkiä. Yksittäisillä kappaleilla kuitenkin suhteellisen nopea.

Menetelmä toimii melko hyvin sukukielillä: raportoitu 4% virhemäärä. Kun lisäksi pyrittiin erikseen tunnistamaan epäilyttävät linjaukset, ja linjaamaan vain parhaat 80% päästiin virhetasoon 0.7

Menetelmä toimii parhaiten 1:1-linjauksilla (2%), mutta hankalammille linjauksille virheprosentit ovat suuria.

Church, 1993: Identtisiin merkkijonoihin perustuva menetelmä

Edelliset menetelmät eivät sovellu kohinaiseen tekstiin (esim. optisen tekstintunnistuksen tuottamaan), jossa saattaa olla roskaa välissä tai kokonaan kadonneita kappaleita. Myös kappale- ja lauserajat ovat vaikeita havaita mm. kadonneiden välimerkkien tai roskan takia.

Tämän menetelmän perustana oleva huomio:

Teksteissä, jotka on kirjoitettu jokseenkin samalla aakkostolla (esim. roomalaiset aakkoset), esiintyy samaatarkoittavia, identtisiä kirjainsekvenssejä kuten erisnimiä tai numeroita.

Sukulaiskielillä, tai läheisessä vuorovaikutuksessa olevilla kielillä voi lisäksi esiintyä muitakin yhteisiä sekvenssejä johtuen yhteisestä kantamuodosta (esim englannin 'superior' ja ranskan 'supérieur') tai lainasanoista.

Lasketaan identtisiä merkki-n-grammeja (n esim. 4). Etsitään n-grammien linjaus joka sisältää mahdollisimman paljon identtisiä n-grammipareja. Lisäksi

n-grammeja voidaan painottaa frekvenssin mukaan.

Menetelmä ei tuota varsinaista lauseiden jyvitystä.

Voi epäonnistua täydellisesti mikäli kielissä ei ole riittävästi yhteisiä merkkijonoja.

Leksikaaliset menetelmät

Tavoitteena on tuottaa aito lausetason 'jyvitys'.

Vaikuttaa selvältä, että tieto sanojen todennäköisistä käännöspareista auttaisi linjausta huomattavasti.

Puhtaasti tilastollisten menetelmien keskeinen ajatus:
vuorotellaan todennäköisen osittaislinjauksen tekemistä sanatasolla ja todennäköisimmän lausetason linjauksen tekemistä.

Apuna käytetään lisäksi oletusta, että toisiaan vastaavat lausejonot eivät luultavasti ole kovin kaukana toisistaan (esim. ristiinmenoja ei ole tai ne eivät ole pitkiä).

Iteraatioita ei yleensä tarvita kovin monta (johtuen yo. rajoituksesta).

Variantteja

Chen, 1993: Sovelletaan yksinkertaista sana-sana-käännösmallia estimoimaan sanaparien käännöstn:t. Lasketaan tällä mallilla maksimaalisen todennäköinen linjaus.

Hyviä puolia: aidon käännösmallin käyttö oletettavasti parantaa tarkkuutta puoliheuristisiin menetelmiin verrattuna. Yksinkertaisen käännösmallin käyttö tekee menetelmästä laskennallisesti tehokkaan.

Huono puoli: mallin yksinkertaisuuden aiheuttamat approksimaatiot voivat aiheuttaa ongelmia kun pitävät huonosti paikkaansa jollekin kieliparille tai tekstiparille.

Menetelmää on sovellettu suurten korpusten (useita miljoonia lauseita) linjaamiseen. Sen virheprosentiksi on estimoitu 0.4% mikä on yhtä hyvä tai parempi kuin muilla menetelmillä saman korpuksen osajoukolla.

Fung ja McKeown, 1994

Estimoidaan pieni kaksikielinen sanakirja, joka antaa todennäköisesti vastaavia sana-käännös-pareja. Käytetään näitä 'ankkureina' linjauksessa, kuten aiemmin käytettiin yhteisiä n-grammeja.

Ei tarvita yhteisiä sanoja tai n-grammeja.

Sanojen linjaus ja kaksikielisten sanakirjojen estimointi

Sanatason linjauksen peruslähestymistapa: vuorotellaan seuraavia askeleita:

1. muodostetaan jokin sanatason linjaus
2. estimoidaan sen perusteella sanaparien käännöstodennäköisyydet

Sovelletaan siis EM-tyyppistä algoritmia.

Kaksikieliseen sanakirjaan hyväksytään (lopulta) vain sanaparit, joista on saatu riittävästi evidenssiä eli esim. riittävän monta näytettä kyseisten sanojen vastaavuudesta.

Voidaan olettaa, että jatkossa sanojen (ja lauseiden) linjauksessa käytetään myös kaksikielisten sanakirjojen sisältämää tietoa.

Suoraviivainen sovellustapa olisi käyttää tällaista sanakirjaa initialisoimaan edellä esitetyn algoritmin sanaparien käännöstodennäköisyydet.

1.3 Konekääntäminen

Kohinainen kanava -mallin eräs sovellus on konekääntäminen (malli esiteltiin puhuttaessa Shannonin informaatioteorian yhteydessä).

Kun halutaan rakentaa malli, jolla käännetään tekstiä englannista e suomeksi s, ajatellaan, että on olemassa kohinainen kanava, johon syötetään tekstiä suomeksi ja se tulee ulos englanniksi. Meidän on siis ainoastaan dekoodattava kohinainen signaali takaisin suomeksi.

Tarvittavaan menetelmään kuuluu kolme osaa:

1. Kielimalli $P(s)$ kertoo suomen lauseiden tn:t,
2. Käännösmalli $P(e|s)$ kertoo lauseiden kääntymistn:t englanniksi,
3. Dekooderi $\hat{s} = \arg\max_s P(s|e)$ laskee todennäköisimmän suomenkielisen lauseen kun tunnetaan englanninkielinen.

Kielimallien estimointia on käsitelty laajasti kurssin muissa osissa.

Käännösmalli

Valitaan tässä käännösmalliksi hyvin yksinkertainen sana-käännösmalli, joka olettaa, että jokainen englannin sana generoituu (kääntyy) 0 tai 1:stä suomen sanasta, kun taas suomen sana voi vastata useaa englannin sanaa. Oletetaan lisäksi, että peräkkäisten sanojen generoituminen (kääntyminen) on toisistaan riippumatonta.

Olkoon s suomenkielinen lause ja e englanninkielinen. Tällöin käännöstn on

$$P(e|s) = \frac{1}{Z} \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{m} P(e_j|s_{a_j}), \tag{5}$$

jossa l ja m ovat sanojen lukumäärät lauseissa s ja e, ja $P(e_j|s_{a_j})$ on tn, jolla sana englannin lauseessa positiossa j generoituu suomen sanasta, joka on positiossa a_j (0 tarkoittaa tyhjää joukkoa). Z on normalisointitekijä.

Sisäkkäiset summaukset summaavat siis yli kaikkien mahdollisten vaihtoehtoisten linjausten, ja kertolasku kertoo yli sanajonon.

Dekooderi

Dekooderin estimointi tapahtuu samoilla periaatteilla ja menetelmillä kuin aiemmin kurssilla on esitetty (ks. luku Markov-malleista).

$$\hat{s} = \arg\max_s P(s|e) = \arg\max_s \frac{P(e|s)P(s)}{P(e)} = \arg\max_s P(e|s)P(s) \quad (6)$$

Lisäksi voidaan huomioida

sanojen sijainti lauseessa: esim. että linjaukset, joissa suomen lauseen alussa oleva sana linjataan englannin lauseen lopussa olevaan sanaa ovat epätodennäk kuin lähempänä toisiaan olevat vastaavuudet.

sanan hedelmällisyys: sanan tn. tuottaa useita sanoja kohdekieleen (jotkut sanat ovat toisia hedelmällisempiä)

Tuloksia

Mallin soveltaminen englannista ranskaksi kääntämiseen Hansard-korpuksella antoi tulokseksi vain 48% oikein käännettyjä lauseita.

Virheissä oli sekä kieliopillisia että semanttisia virheitä.

Ongelmia

Näin yksinkertaisella lähestymistavalla on monia ilmiselviä ongelmia, jotka johtuvat mm. siitä, että käännetään suoraan sanatasolla.

Joitain menetelmälle ominaisia ongelmia

- herkkyys opetusdatalle: pienet muutokset opetusdatan (tai testidatan) valinnassa aiheuttavat suuria muutoksia tulosprosentteihin. Vastaavuuden testi- ja opetusdatan välillä on oltava hyvin suuri, jotta tällainen sanatason käännösmalli toimisi hyvin.
- Tehokkuus: raskas pitkille lauseille
- Datan harvuus (riittämättömyys)
- Jos kielimalli on lokaali (esim. n-grammimalli), ei auta vaikka käännösmalli osaisi tuottaa käännöksiä hyödyntäen pitkän matkan riippuvuuksia. Eri mallien tekemien oletusten pitäisi olla konsistentteja.