Statistical and Adaptive Natural Language Processing

T-61.5020 (5 cr) L, Spring 2007

Tenth lecture

Statistical Machine Translation

Lecturer: **Mathias Creutz**

Slides: Krista Lagus, Philipp Köhn, Mathias Creutz, Timo Honkela

10. Statistical Machine Translation

Lecture based on:

- Chapter 13 in Manning & Schütze
- Chapter 21 in Jurafsky & Martin: Speech and Language Processing (An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition)
- Article on the IBM model: Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra and Robert L. Mercer (1993). The Mathematics of Statistical Machine Translation. *Computational Linguistics* 19 (2), pp. 263–312.
- Slides by Philipp Köhn (Koehn), Lecturer (Assistant Professor) at the University of Edinburgh.

10.1 On (Machine) Translation in General

- Automaattinen kielenkääntäminen on eräs pitkäaikaisimmista kieliteknologian tavoitteista.
- Konekääntäminen (machine translation) on kuitenkin hyvin vaikea ongelma.
- Nykyisten konekäännösohjelmien tulos toimii lähinnä raakakäännöksenä, joka voi nopeuttaa aidon kielenkääntäjän työtä, mutta ei välttämättä 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, *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 (analysis, transfer, generation).

- *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 a 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.

  By the way, this is a classical example of how the *manner* and *direction* of motion are expressed differently in different languages: In English the verb indicates the manner, whereas the satellite (added words) indicates the direction: *crawl out, float off, jump down, walk over to,*

run after. In Spanish, the opposite is true: *salir flotando* (exit floating = float out), *acercarse corriendo* (approach running = run towards), *alcanzar andando* (reach walking = reach by foot = walk all the way to).

Curiously enough, there also exist languages, where the verb describes *the shape of the one who moves or is moved*: something like *to slither* (a snake?).

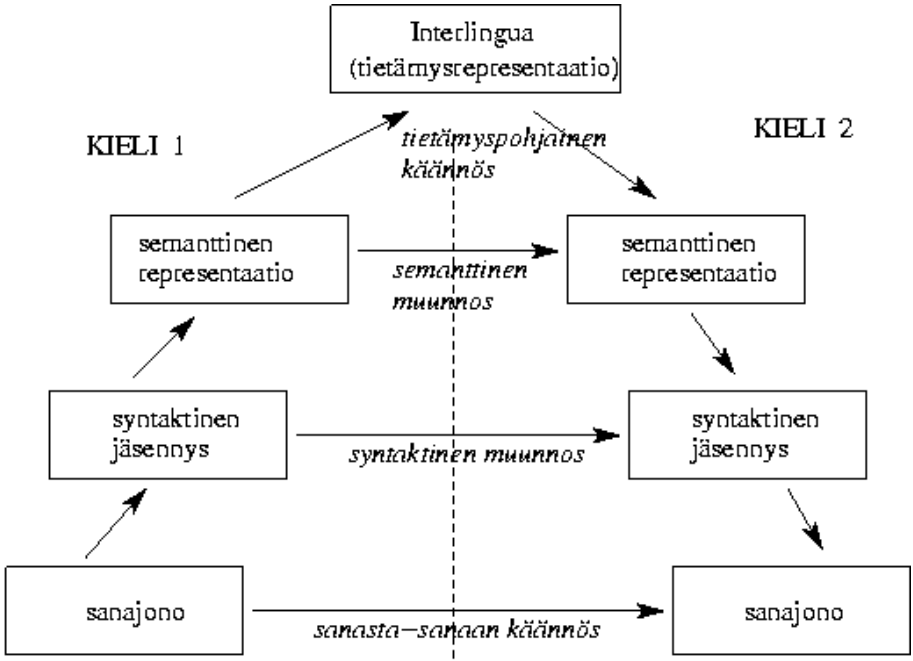
- *Interlingua* – keinotekoinen yleinen (kieliriippumaton) välikieli tai tietämysrepresentaatio. Käännetään lähtökielestä interlingualle ja interlinguasta mille tahansa kohdekielelle.

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

Seuraavan sivun kuvassa on näytetty konekäännösjärjestelmän vaihtoehtoiset toteutustavat.

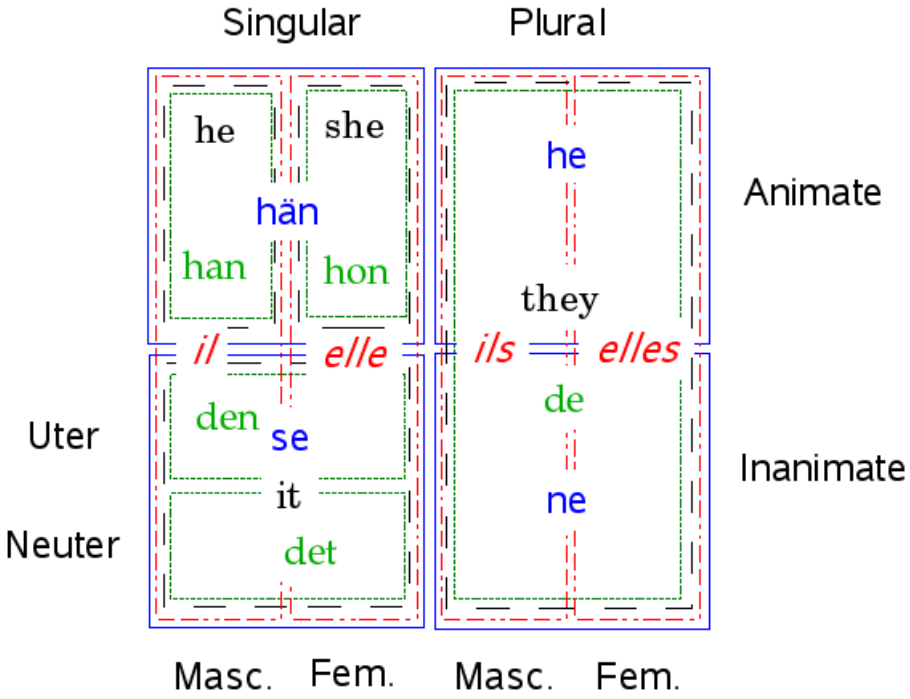
Tilastollisen kielenkäsittelyn menetelmiä voidaan käyttää järjestelmän komponentteina minkä tahansa nuolen kohdalla (esim. jäsentäminen, disambiguointi jne).

Konekääntimet voivat myös olla kombinaatioita symbolisista ja tilastollisista komponenteista.


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

Semantic differences across languages

- Distinctions are made in different ways in different languages. That is, the "semantic space" is divided differently. For instance, if we are to translate the English color word *blue* into Russian, we have to know whether it is dark blue (*sinij*) or light blue (*goluboj*). (There is no word for "just blue" in Russian, but if we have to choose without any additional knowledge, we would probably take *sinij*.)
- Also compare some third-person pronouns in four languages. Distinctions are made based on number (singular, plural), animacy (animate, inanimate), and gender (masculine, feminine, uter, neuter):
  - Finnish: *hän, se, he, ne* (no gender distinction)
  - Swedish: *han, hon, den, det, de* (animacy and gender distinction only in singular)
  - English: *he, she, it, they* (animacy and gender only in singular)
  - French: *il, elle, ils, elles* (no animacy distinction)


Singular
Plural
he
she
han
hon
hän
he
they
Animate
il
elle
ils
elles
Uter
den
se
de
Inanimate
it
ne
Neuter
det
Masc.
Fem.
Masc.
Fem.

- Cultural and linguistic differences: **Sapir-Whorf hypothesis**: language constrains thought – the language you speak may affect the way you think.
  Some things are left ambiguous in one language, whereas they have to be fixed in another. For instance, in Russian you have to decide between dark and light blue, and in many languages you have to decide between "he" and "she". (Does this mean that Finnish speakers are more likely to think of persons gender-neutrally? Does it mean that French speakers associate masculine and feminine properties with inanimate objects? Does a plant (*une plante*) have female features, whereas a tree (*un arbre*) has masculine ones?)

- To the extent that the Sapir-Whorf hypothesis is true, there can be no perfect translation, since speakers of the source and target languages necessarily have different conceptual systems.

- Couldn't interlingua be the solution: an artifical language with very fine-grained semantic categories and distinctions?

- Problem 1: How construct such an intrinsically complicated system?
- Problem 2: If two languages are rather closely related, they may share a number of constructs and ambiguities. If we were to translate the Swedish sentence "De kom för sent." into English, we could actually do almost with a word-to-word translation: "They came too late.". It does not matter whether "they" are animate or inanimate ("he ihmiset" or "ne taksit") or whether they came walking, swimming, or driving. It does not seem optimal to require a parser to perform deeper analysis and more disambiguation than necessary.

- Before moving on to statistical methods, we shall take a brief look at two examples of **rule-based direct translation**, which does not make use of complex structures and representations (no interlingua). The input is treated as a string of words (or morphemes), and various operations are performed directly on it.

Japanese-to-English Translation

Stage	Action
1.	morphological analysis
2.	lexical transfer of content words
3.	various work relating to prepositions
4.	SVO rearrangements
5.	miscellany
6.	morphological generation

Input:	watashihatsukuenouenopenwojonniageta.
After stage 1:	watashi ha tsukue no ue no pen wo jon ni ageru PAST.
After stage 2:	*I* ha *desk* no ue no *pen* wo *John* ni *give* PAST.
After stage 3:	I ha pen on desk wo John to give PAST.
After stage 4:	I give PAST pen on desk John to.
After stage 5:	I give PAST the pen on the desk to John.
After stage 6:	I gave the pen on the desk to John.

Direct translation of 'much' and 'many' into Russian

```
if preceding word is how /* how much, how many */
      return skol'ko
else if preceding word is as /* as much, as many */
      return stol'ko zhe
else if word is much
      if preceding word is very /* very much */
            return nil (not translated)
      else if following word is a noun /* much people, food */
            return mnogo
else /* word is many */
      if preceding word is a preposition and following word is a noun
            return mnogii
else return mnogo
```

Adapted from Hutchin's (1986) discussion of Panov (1960).

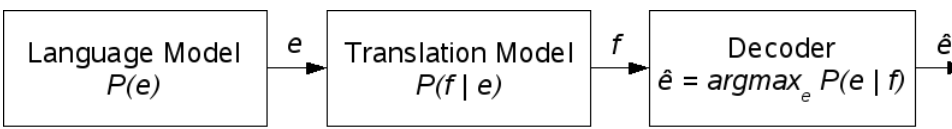
Imagine that you'd have to debug such a system with all its word-specific rules...

10.2 Statistical Approach

- In 1949, Warren Weaver suggested applying statistical and cryptanalytic techniques from the field of communication theory to the problem of using computers to translate text from one natural language to another.

- However, computers at that time were far too inefficient, and the availability of language data (text) in digital form was very limited.

- The idea of the **noisy channel** model: The language model generates an English sentence e. The translation model transmits e "noisily" as the foreign sentence f. The decoder finds the English sentence $\hat{e}$ which is most likely to have given rise to f.



- In the examples, we usually translate from a foreign language f into English e. (The Americans want to figure out what is written or spoken in Russian, Chinese, Arabic...) In the first publications in the field (the so-called IBM model), f referred to French, but to think of f as any foreign language is more general.
- Using Bayes' rule, or the noisy channel metaphor, we obtain:

$$P(e|f) = \frac{P(e)P(f|e)}{P(f)}. \tag{1}$$

  Since the denominator is independent of e, finding $\hat{e}$ is the same as finding e so as to make $P(e)P(f|e)$ as large as possible:

$$\hat{e} = \arg\max_e P(e)P(f|e). \tag{2}$$

- This can be interpreted as maximizing the **fluency** of the English sentence $P(e)$ as well as the **faithfulness** of the translation between English and the foreign language $P(f|e)$:

$$\text{best translation } \hat{e} = \arg\max_e \ \text{fluency}(e) \cdot \text{faithfulness}(f|e). \tag{3}$$

- The language model probability (or measure of fluency) $P(e)$ is typically decomposed into a product of n-gram probabilities (see Lecture 9).
- The translation model (or measure of faithfulness) $P(f|e)$ is typically decomposed into a product of word-to-word, or phrase-to-phrase, translation probabilities. For instance, $P(\textit{Angleterre}|\textit{England})$ should be high, whereas $P(\textit{Finlande}|\textit{England})$ should be low.
- Maybe strange to think of a human translator that would divide the task into first (1) enumerating a large number of fluent English sentences, and then (2) choosing one, where the words translated into French would match the French input sentence well.
- The IBM model also comprises **fertility** and **distortion** probabilities. We will get back to them shortly.
- The success of statistical machine translation depends heavily on the quality of the **text alignment** that is produced.

10.3 Text Alignment

- Tekstinkohdistuksella (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 (e.g., Europarl, Canadian and Hong Kong Hansards, KOTUS Finnish-Swedish parallel corpus).
- 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).

- Tekstinkohdistuksessa on yleensä kaksi vaihetta:

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

Lauseiden ja kappaleiden kohdistus

Yleensä lauseiden kohdistus on välttämätön ensimmäinen askel monikielisen korpuksen tuottamisessa.

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

- Sananmerkitysten disambiguointi: sanan eri merkityksiä voidaan ryhmitellä sen saamien eri käännösvastineiden perusteella. For instance, the Finnish word *kuusi* can be translated as *six* or *spruce* (or *your moon*).
- 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 (kohdistetun tekstin) toisenkielinen lausejono. Kumpi tahansa jono voi olla myös tyhjä. Jokainen lause kuuluu täsmälleen yhteen jyvään.

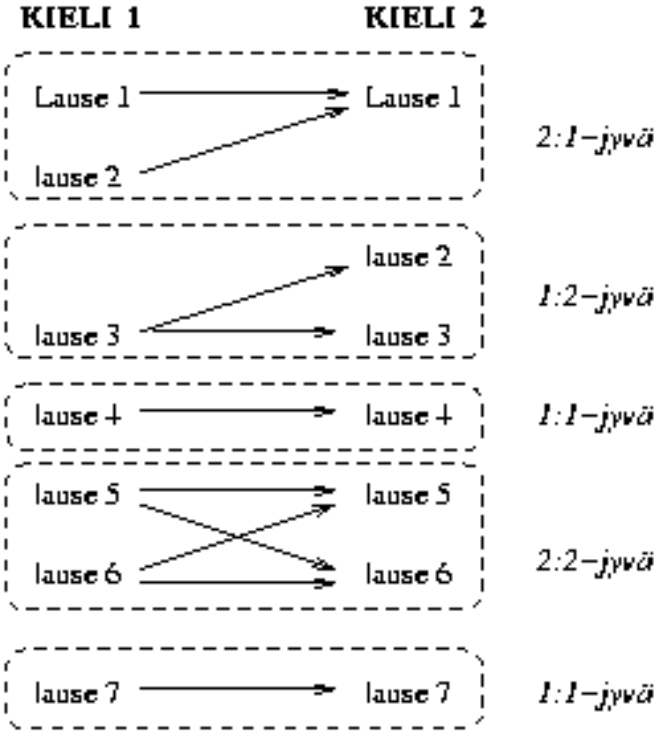
Jyvitys on kuvaus, jossa tekstit on jaettu osiin ja kerrottu, mitä kielen 1 osaa mikäkin kielen 2 osa vastaa.

Lauseiden kohdistus 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

Example of a 2:2 Alignment

The sentence divisions in the English and French texts are different:

- *English:* According to our survey, 1988 sales of mineral water and soft drinks were much higher than in 1987, reflecting the growing popularity of these products.
  Cola drink manufacturers in particular achieved above average growth rates.

- *French:* Quant aux eaux minérales et aux limonades, elles rencontrent toujours plus d'adeptes.
  En effet notre sondage fait ressortir des ventes nettement supérieures à celles de 1987, pour les boissons à base de cola notamment.

- *French-to-English literal translation:* With regard to the mineral waters and the lemonades, they encounter still more users.
  Indeed our survey makes stand out the sales clearly superior to those in 1987 for cola-based drinks especially.

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.

Tekstinkohdistuksen 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 F jono lauseita $F = (f_1, \ldots, f_I)$ ja kielen 2 teksti E samoin $E = (e_1, \ldots, e_J)$ (F = foreign, E = English)

Tekstinpätkien pituuksiin perustuvat menetelmät

Useat varhaiset tekstinkohdistusmenetelmät ovat tätä tyyppiä.

Etsitään kohdistus (alignment) A, jolla on suurin tn:

$$\arg\max_A P(A|F,E) = \arg\max_A P(A,F,E) \tag{4}$$

(todennäköisin kohdistus voidaan etsiä mm. dynaamisella ohjelmoinnilla).

Useat menetelmät jakavat kohdistetun tekstin jonoksi jyviä (beads) $(B_1,\ldots,B_K)$ ja approksimoivat koko kohdistetun tekstin tn:ää olettamalla, että jyvän tn ei riipu ympäristön lauseista tai niiden jyvityksestä, vaan ainoastaan kyseisen jyvän sisältämistä lauseista:

$$P(A,F,E) = \prod_{k=1}^{K} P(B_k) \tag{5}$$

Jyvän todennäköisyyden laskenta

Gale & Church, 1991, 1993:

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

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

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

Olkoon $D(i,j)$ etsitty pienimmän kustannuksen kohdistus lauseiden $f_1, \ldots, f_i$ ja $e_1, \ldots, e_j$ välillä.

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

Rekursio:

$$
\begin{aligned}
D(i,j) = \min \quad & D(i,j-1) && + cost(0:1 \text{ kohdistus } 0, e_j) \\
& D(i-1,j) && + cost(1:0 \text{ kohdistus } f_i, 0) \\
& D(i-1,j-1) && + cost(1:1 \text{ kohdistus } f_i, e_j) \\
& D(i-1,j-2) && + cost(1:2 \text{ kohdistus } f_i, e_{j-1}, e_j) \\
& D(i-2,j-1) && + cost(2:1 \text{ kohdistus } f_{i-1}, f_i, e_j) \\
& D(i-2,j-2) && + cost(2:2 \text{ kohdistus } f_{i-1}, f_i, e_{j-1}, e_j)
\end{aligned}
$$

Kunkin tyyppisen kohdistuksen (jyvän) kustannus lasketaan seuraavasti:

Oletetaan malli: yksi kielen L_1 merkki generoi satunnaisen määrän merkkejä kieleen L_2. Oletetaan generoitujen merkkien määrän noudattavan gaussista tn-jakaumaa. Jakauman keskiarvo μ ja varianssi σ^2 estimoidaan suurista rinnakkaiskorpuksista (saksa/englanti-parille estimoitiin $\mu = 1.1$ koko korpuksesta, ranska/englanti-parille 1.06.)

Kustannuksena voidaan käyttää tekstinpätkien etäisyyden negatiivista log-

likelihoodia mallissa:

$$cost(l_1, l_2) = -\log P(\alpha \text{ kohdistus } |\delta(l_1, l_2, \mu, \sigma^2) \quad (6)$$

jossa α on jokin sallituista kohdistustyypeistä ja σ mittaa merkien määrän keskiarvoa ja varianssin eroa niiden korpus-estimaateista: $\delta(l_1, l_2, \mu, \sigma^2) = (l_2 - l_1\mu)/\sqrt{l_1\sigma^2}$.

Tarvittavat todennäköisyydet estimoidaan soveltamalla Bayesin kaavaa

$$P(\alpha|\delta) = P(\alpha)P(\delta|\alpha) \quad (7)$$

Tällöin siis 1:1-kohdistuksen suuri prioritodennäköisyys ($P(\alpha = 1:1) = 90\,\%$) aiheuttaa sen kohdistuksen suosimista.

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 kohdistukset, ja linjaamaan vain parhaat 80% päästiin virhetasoon 0.7

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

Brown et al 1991:

Edellisen menetelmän variantissa lasketaan lauseiden pituuksia sanojen lukumääränä merkkien lukumäärän sijaan. On argumentoitu että tämä on huonompi tapa koska sanojen lukumäärissä on enemmän varianssia kuin merkkien määrissä.

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 kohdistus 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 kohdistusta huomattavasti.

Puhtaasti tilastollisten menetelmien keskeinen ajatus:
vuorotellaan todennäköisen osittaiskohdistuksen tekemistä sanatasolla ja todennäköisimmän lausetason kohdistuksen 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).

10.4 Word Alignment

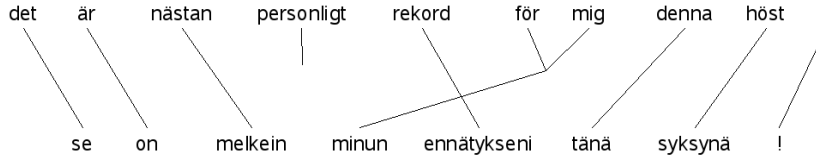
- In the alignment of entire sentences and sections, we did not identify cross-alignments. If there were differences in the order in which the message was conveyed in the two languages, we created large enough beads that comprised multiple sentences on both sides. In this way, we didn't have to rearrange the order of the sentences in either language, while each bead still contained approximately the same thing in both languages.

- The sentence alignment was just a first step to facilitate a complete word-level alignment. In the word-level alignment, we do take into account the reordering (called *distortion*) and *fertility* of the words.

- Distortion means that word order differs across languages.

- The fertility of a word in one source language with respect to another target language measures how many words in the target language the word in the source language is translated to on average.

- For instance,



Personligt was not aligned at all, and the two words *för mig* were aligned with one word *minun* (and the morpheme *-ni* if we analyze the words into parts).

- Sanatason kohdistuksen peruslähestymistapa: vuorotellaan seuraavia askeleita:

  1. muodostetaan jokin sanatason kohdistus
  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.

- The translation probability of a sentence is then obtained as: Olkoon f vieraskielinen lause ja e englanninkielinen. Tällöin käännöstodennäköisyys on

$$P(f|e) = \frac{1}{Z} \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{m} P(f_j|e_{a_j}), \tag{8}$$

  jossa l ja m ovat sanojen lukumäärät lauseissa e ja f, ja $P(f_j|e_{a_j})$ on todennäköisyys, jolla sana vieraskielisessä lauseessa positiossa j generoituu englannin sanasta, joka on positiossa a_j (0 tarkoittaa tyhjää joukkoa). Z on normalisointitekijä.

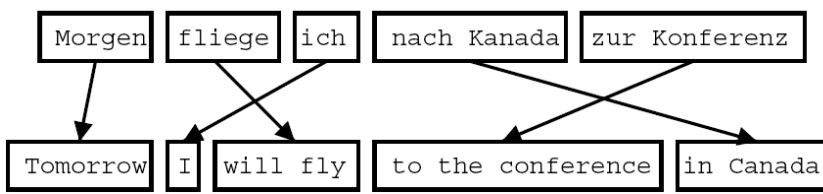
  Sisäkkäiset summaukset summaavat siis yli kaikkien mahdollisten vaihtoehtoisten kohdistusten, ja kertolasku kertoo yli sanajonon.

- The word-level translation probability can be constructed so as to take into account distortion and fertility probabilities.

10.5 Phrase Alignment

- Problems with word-to-word translation:
  - "Cut-and-paste" translation (no syntax or semantics): it is probable that when words are "cut" from one context and "pasted" into another context mistakes occur, despite the language model.
  - The distortion (reordering) probability typically penalizes more, if several words have to be reordered. However, usually larger multi-word chunks (subphrases) need to be moved.

- Example:



- Phrase-to-phrase translation is an alternative to the IBM word-to-word model.
- Although we still rely on the "cut-and-paste" philosophy, we deal with larger chunks, so there are fewer "seams" between chunks combined in a new way. The word sequence within a phrase has been attested before in real texts, so it should be more or less correct. Phrases can also capture non-compositional word sequences, such as *it's anyone's guess* = *on mahdoton tietää*. In short, better use is made of the **local context**.
- The more data, the longer phrases can be learned.

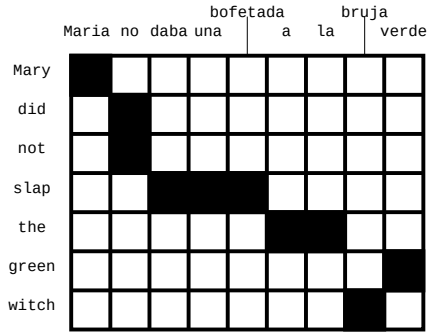
Phrase translation table

- Phrase translations for *den Vorschlag*

English	ϕ(e|f)	English	ϕ(e|f)
the proposal	0.6227	the suggestions	0.0114
's proposal	0.1068	the proposed	0.0114
a proposal	0.0341	the motion	0.0091
the idea	0.0250	the idea of	0.0091
this proposal	0.0227	the proposal ,	0.0068
proposal	0.0205	its proposal	0.0068
of the proposal	0.0159	it	0.0068
the proposals	0.0159	...	...

How to learn the phrase translation table?

- Start with the *word alignment*:



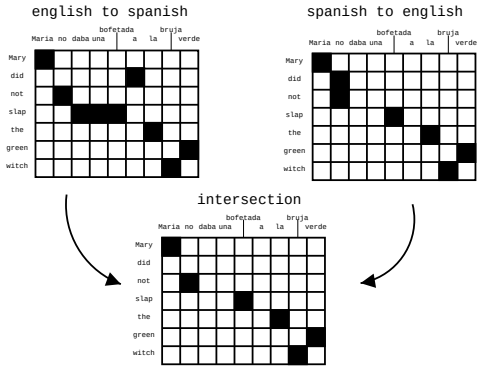
- Collect all phrase pairs that are **consistent** with the word alignment

Word alignment with IBM models

- IBM Models create a *many-to-one* mapping
  - words are aligned using an **alignment function**
  - a function may return the same value for different input (one-to-many mapping)
  - a function can not return multiple values for one input (*no many-to-one* mapping)

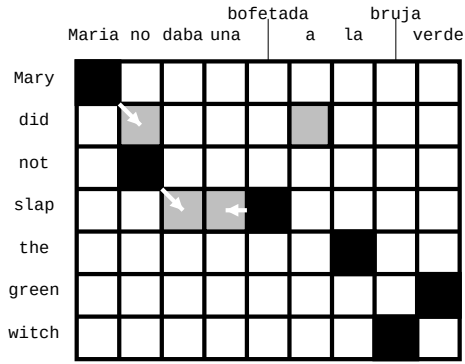
- But we need *many-to-many* mappings

Symmetrizing word alignments



- *Intersection* of GIZA++ bidirectional alignments

Symmetrizing word alignments



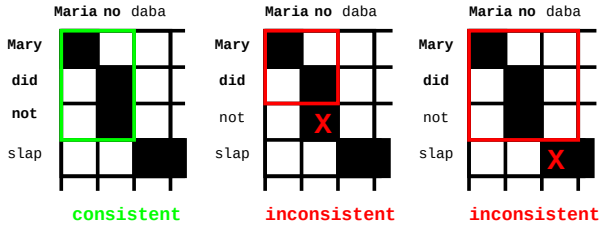
- *Grow* additional alignment points [Och and Ney, CompLing2003]

Growing heuristic

```
GROW-DIAG-FINAL(e2f,f2e):
  neighboring = ((-1,0),(0,-1),(1,0),(0,1),(-1,-1),(-1,1),(1,-1),(1,1))
  alignment = intersect(e2f,f2e);
  GROW-DIAG(); FINAL(e2f); FINAL(f2e);

GROW-DIAG():
  iterate until no new points added
    for english word e = 0 ... en
      for foreign word f = 0 ... fn
        if ( e aligned with f )
          for each neighboring point ( e-new, f-new ):
            if ( ( e-new not aligned and f-new not aligned ) and
                 ( e-new, f-new ) in union( e2f, f2e ) )
              add alignment point ( e-new, f-new )
FINAL(a):
  for english word e-new = 0 ... en
    for foreign word f-new = 0 ... fn
      if ( ( e-new not aligned or f-new not aligned ) and
           ( e-new, f-new ) in alignment a )
        add alignment point ( e-new, f-new )
```

Consistent with word alignment

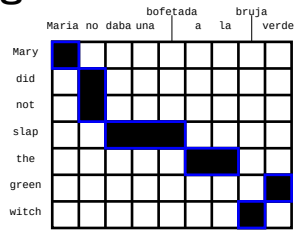


- **Consistent with the word alignment** :=

  phrase alignment has to *contain all alignment points* for all covered words

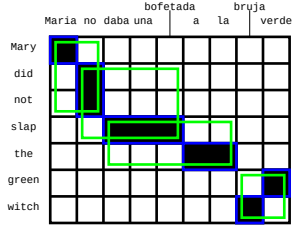
$$(\overline{e}, \overline{f}) \in BP \Leftrightarrow \quad \forall e_i \in \overline{e} : (e_i, f_j) \in A \rightarrow f_j \in \overline{f}$$
$$\text{AND} \quad \forall f_j \in \overline{f} : (e_i, f_j) \in A \rightarrow e_i \in \overline{e}$$

Word alignment induced phrases



(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green)

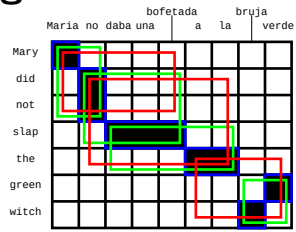
Word alignment induced phrases



(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green),

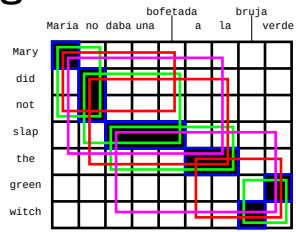
(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the), (bruja verde, green witch)

Word alignment induced phrases



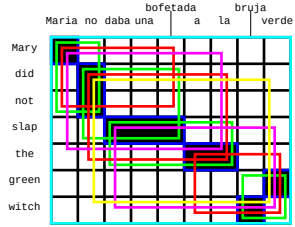
(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green),
(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the),
(bruja verde, green witch), (Maria no daba una bofetada, Mary did not slap),
(no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch)

Word alignment induced phrases



(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green),
(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the),
(bruja verde, green witch), (Maria no daba una bofetada, Mary did not slap),
(no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch),
(Maria no daba una bofetada a la, Mary did not slap the),
(daba una bofetada a la bruja verde, slap the green witch)

Word alignment induced phrases (5)



(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green),
(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the),
(bruja verde, green witch), (Maria no daba una bofetada, Mary did not slap),
(no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch),
(Maria no daba una bofetada a la, Mary did not slap the), (daba una bofetada a la bruja verde,
slap the green witch), (no daba una bofetada a la bruja verde, did not slap the green witch),
(Maria no daba una bofetada a la bruja verde, Mary did not slap the green witch)

Probability distribution of phrase pairs

- We need a **probability distribution** $\phi(\overline{f}|\overline{e})$ over the collected phrase pairs

$\Rightarrow$ Possible *choices*

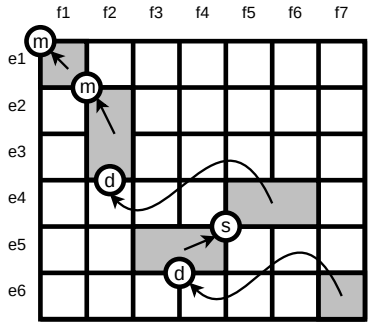
- *relative frequency* of collected phrases: $\phi(\overline{f}|\overline{e}) = \frac{\text{count}(\overline{f},\overline{e})}{\sum_{\overline{f}} \text{count}(\overline{f},\overline{e})}$
- or, conversely $\phi(\overline{e}|\overline{f})$
- use *lexical translation probabilities*

Reordering

- *Monotone* translation
  - do not allow any reordering

  → worse translations
- *Limiting* reordering (to movement over max. number of words) helps
- *Distance-based* reordering cost
  - moving a foreign phrase over n words: cost ω^n
- *Lexicalized* reordering model

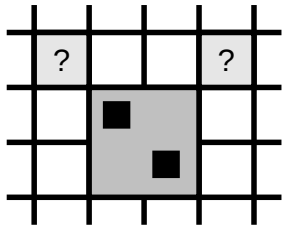
Lexicalized reordering models



[from Koehn et al., 2005, IWSLT]

- Three **orientation** types: **monotone**, **swap**, **discontinuous**
- Probability $p(swap|e, f)$ depends on foreign (and English) *phrase* involved

Learning lexicalized reordering models



[from Koehn et al., 2005, IWSLT]

- Orientation type is *learned during phrase extractions*
- *Alignment point* to the *top left* (monotone) or *top right* (swap)?
- For more, see [Tillmann, 2003] or [Koehn et al., 2005]

10.6 Evaluation

- **BLEU** (Bilingual Evaluation Understudy) is a method for evaluating the quality of text which has been translated from one natural language to another using machine translation. BLEU was one of the first software metrics to report high correlation with human judgments of quality.

- The metric calculates scores for individual segments, generally sentences, and then averages these scores over the whole corpus in order to reach a final score.

- The metric works by measuring the n-gram (1, 2, 3, and 4-gram) co-occurrence between a given translation and the set of reference translations and then taking the weighted geometric mean. BLEU is specifically designed to approximate human judgment on a corpus level and performs badly if used to evaluate the quality of isolated sentences.

Philipp Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. MT Summit 2005.

Source Language	Target Language										
	da	de	el	en	es	fr	fi	it	nl	pt	sv
da	-	18.4	21.1	28.5	26.4	28.7	14.2	22.2	21.4	24.3	28.3
de	22.3	-	20.7	25.3	25.4	27.7	11.8	21.3	23.4	23.2	20.5
el	22.7	17.4	-	27.2	**31.2**	**32.1**	11.4	26.8	20.0	27.6	21.2
en	25.2	17.6	23.2	-	**30.1**	**31.1**	13.0	25.3	21.0	27.1	24.8
es	24.1	18.2	28.3	**30.5**	-	**40.2**	12.5	**32.3**	21.4	**35.9**	23.9
fr	23.7	18.5	26.1	**30.0**	**38.4**	-	12.6	**32.4**	21.1	**35.3**	22.6
fi	20.0	14.5	18.2	21.8	21.1	22.4	-	18.3	17.0	19.1	18.8
it	21.4	16.9	24.8	27.8	**34.0**	**36.0**	11.0	-	20.0	**31.2**	20.2
nl	20.5	18.3	17.4	23.0	22.9	24.6	10.3	20.0	-	20.7	19.0
pt	23.2	18.2	26.4	**30.1**	**37.9**	**39.0**	11.9	**32.0**	20.2	-	21.9
sv	**30.3**	18.9	22.8	**30.2**	28.6	29.7	15.3	23.9	21.9	25.9	-

Table 2: BLEU scores for the 110 translation systems trained on the Europarl corpus

Examples of Phrase-Based Translation (Europarl Swedish-to-Finnish)

The open-source statistical machine translation system **Moses** has been used (`http://www.statmt.org/moses/`). Moses was trained on text data in which the words had been split into morphs by Morfessor The training set contained circa 900,000 sentences, or 20 million words (including punctuation marks).

The borders of the phrases used are marked using a vertical bar |. Morph boundaries are not marked:

- **Source 1:** det är nästan personligt rekord för mig denna höst !
- **Translation 1:** se on melkein | henkilökohtainen | ennätys | minulle | tämän | vuoden syksyllä | !
- **Reference 1:** se on melkein minun ennätykseni tänä syksynä !

- **Source 2:** det är fullständigt utan proportioner och hjälper inte till i fredsprocessen på något sätt .
- **Translation 2:** se on täysin | ilman | suhteelli|suudentaju | ja auttaa | rauhanprosessissa | ei | millään | tavalla .
- **Reference 2:** tämä on täysin suhteetonta eikä se edistä rauhanprosessia millään tavoin .

- **Source 3:** jag går in på denna punkt därför att den är mycket intressant .
- **Translation 3:** en | käsittele | tätä kohtaa | , koska se | on hyvin mielenkiintoinen .
- **Reference 3:** puutun tähän kohtaan , koska se on hyvin mielenkiintoinen .

- **Source 4:** vad konkurrensen anbelangar så är marknaden avgörande för utvecklingen i kusthamnarna .
- **Translation 4:** mitä | tulee | niin | kilpailu|t | markkinat ovat | ratkaisevan tärkeitä | kehitykse|n | merisatamiin | .
- **Reference 4:** mitä kilpailuun tulee , markkinat vaikuttavat ratkaisevasti merisatamien kehitykseen .

- **Source 5:** denna prioritering är emellertid skadlig för miljön och innebär ett socialt slöseri .
- **Translation 5:** tämän | ensisijaisena tavoitteena on | kuitenkin | vahingoittaa | ympäristöä ja aiheuttaa | yhteiskunnallista | tuhlausta .
- **Reference 5:** tällainen suosiminen on kuitenkin ekologisesti vahingollista ja sosiaalisesti epäonnistunutta .

Some Weaknesses of the System

- No modeling of syntax or semantics.
- Herkkyys opetusdatalle: pienet muutokset opetusdatan (tai testidatan) valinnassa aiheuttavat suuria muutoksia tulosprosentteihin. Vastaavuuden testi- ja opetusdatan välillä olisi oltava hyvin suuri, jotta tällainen sanatason käännösmalli toimisi hyvin.
- Tehokkuus: raskas pitkille lauseille.
- Datan harvuus (riittämättömyys). Harvinaisten sanojen osalta estimaatit ovat huonoja (lue:melko satunnaisia).
- Morfologisesti rikkaissa kielissä harvan datan ongelma korostuu, ellei sanoja pilkota tms.
- 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.